Universität der Bundeswehr München

Generische Modellierung von zweistufigen Transportketten zur Planung und Optimierung von Verteilzentren in komplexen Systemen

Dipl.-Ing. Peter Hillmann

Vollständiger Abdruck der von der Fakultät für Informatik der Universität der Bundeswehr München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften
(Dr. rer. nat.)

genehmigten Dissertation.

Gutachter/Gutachterin:

Prof. Dr. Gabi Dreo Rodosek
Prof. Dr. Oliver Rose

Die Dissertation wurde am 24.11.2017 bei der Universität der Bundeswehr München eingereicht und durch die Fakultät für Informatik am 16.02.2018 angenommen. Die mündliche Prüfung fand am 14.03.2018 statt.

Prüfungskommission:

Vorsitzender:	Prof. Dr. Peter Hertling
1. Gutachter:	Prof. Dr. Gabi Dreo Rodosek
2. Gutachter:	Prof. Dr. Oliver Rose
1. Prüfer:	Prof. Dr. Uwe Borghoff
2. Prüfer:	Prof. Dr. Wolfgang Hommel
3. Prüfer:	Prof. Dr.-Ing. Andreas Karcher

Kurzfassung

In Transportsystemen werden zentral gelegene Zwischenspeicher zur Reduktion der Transportwege und zur Bündelung von Güterflüssen eingesetzt. Die Bestimmung bestmöglicher Standorte für solche Zentren und die angepasste Zuweisung zur Belieferung sind für einen optimierten Ablauf erforderlich. Diese Herausforderung ist bei verteilten Systemen in der Logistik als auch bei Rechnernetzen wiederzufinden. Im Bereich der Logistik entspricht die Errichtung zentraler Stationen für effiziente Transportketten dem Facility Location Problem. Bei Rechnernetzen ist dies vergleichbar mit der Platzierung von Servern zur Verbesserung des Datentransfers über Content Delivery Networks. Die Dissertation stellt für diese Aufgabenstellungen ein generisches Modell vor, das durch Parametrisierung flexibel für unterschiedliche Domänen adaptierbar ist. Mittels verschiedener Algorithmen erfolgt eine teilproblemübergreifende, multikriterielle Optimierung, sodass eine angepasste und ganzheitliche Lösung erzeugt wird. Über eine Simulation des entstandenen Modells findet eine Analyse und Bewertung des operativen Systemverhaltens statt.

Abstract

In supply chains, central storages are used to reduce transport distances and to combine flows of goods. The selection of appropriate locations and the suited assignment for delivery are significant challenges for an optimized procedure. This characteristic of distributed systems occurs in logistics as well as in computer networks. In the domain of logistics, the establishment of central stations for efficient transport chains corresponds to the facility location problem. In computer networks, this is similar to the placement of servers to improve data transfer via content delivery networks. The PhD thesis presents a generic model for these tasks, which is flexibly adaptable for different domains by configuration. Furthermore, a strategic approach is devised to obtain suitable solutions. Therefore, an overall optimization is carried out, resulting in a tailored and integrated solution with respect to multiple objectives. The operational system behavior are analyzed with a corresponding simulation of the resulting model.

Vorwort

Schon als kleines Kind löste ich gern Rätsel und Knobelaufgaben. Ich erinnere mich noch sehr gut an eine logistische Problemstellung, welche mir meine Eltern gestellt hatten. Bei dieser Flussüberquerungsaufgabe kommt ein Bauer mit einem Wolf, einem Schaf sowie einem Kohlkopf unterwegs an einen Fluss und muss diesen überqueren. Die zur Verfügung stehende Ressource ist ein Boot, wobei der Bauer immer nur eines der Tiere oder den Kohlkopf mitnehmen kann. Als Nebenbedingung gilt, dass der Wolf naturgemäß nicht mit dem Schaf alleine zu lassen ist, ebenso wie das Schaf mit dem Kohlkopf, da der eine den anderen fressen würde. Wie muss der Bauer vorgehen, um mit möglichst wenigen Transportfahrten den Fluss ohne Verlust zu überqueren?

Auch heute noch bin ich motiviert, mich mit wesentlich schwierigeren Aufgaben und Problemstellungen komplexer Systeme zu beschäftigen. Mit Beginn des Schreibens dieser Arbeit war mir klar, dass die Bearbeitung des Themas nicht einfach wird. Aber da dachte ich an den Spruch: „Per aspera ad astra – über raue Pfade gelangt man zu den Sternen.“

Danksagung

Meine Tätigkeit an der Universität der Bundeswehr München ist eine fachliche und persönliche Bereicherung sowie Herausforderung, die mich für die weitere berufliche Laufbahn prägt. Ich blicke zufrieden und dankbar auf eine außergewöhnliche Promotionszeit zurück und freue mich auf eine spannende Zukunft am Forschungsinstitut Cyber Defence (CODE) der Universität der Bundeswehr München.

Frau Prof. Dr. Gabi Dreo und Herrn Prof. Dr. Oliver Rose danke ich herzlich für die Betreuung der Arbeit und die vielen Hinweise, die zur Weiterentwicklung der Dissertation beitrugen. Dieser Dank gilt auch Herrn Dr. Tobias Uhlig, Herrn Falk Stefan Pappert und Herrn Lars Stiemert.

Nützlich für meine Arbeit waren auch die Konferenzen im Ausland, durch die ich weitere Anregungen erhielt.

Danken möchte ich auch allen derzeitigen und ehemaligen Team-Mitgliedern von den beiden Professoren, die stets wichtige Diskussionspartner waren.

Großer Dank gilt auch meinen Eltern, der Familie und vielen Freunden, die für die notwendige Ablenkung in den richtigen Momenten gesorgt haben, mich immer motivierten und mir in schwierigen Situationen zur Seite standen.

München, im November 2017

Peter Hillmann

Peter Hillmann

Inhaltsverzeichnis

1 Einleitung **1**
1.1 Einordnung der Arbeit . . . 6
1.2 Anwendungsgebiet und Motivation . . . 8
1.2.1 Transportkette und -system . . . 8
1.2.2 Motivation zum Einsatz von Verteilzentren . . . 9
1.2.3 Anwendungsgebiet der Logistik . . . 10
1.2.4 Anwendungsgebiet der Rechnernetze . . . 12
1.2.5 Motivation für generischen Ansatz . . . 14
1.2.6 Definition von Begriffen . . . 16
1.3 Wissenschaftliche Fragestellungen . . . 18
1.4 Zielstellung der Arbeit . . . 20
1.5 Aufbau der Dissertation . . . 21

2 Problemstellung **23**
2.1 Generische Modellierung . . . 23
2.1.1 Infrastruktur und Vernetzung der Teilnehmer einer Transportkette . . . 24
2.1.2 Standortmöglichkeiten für Zentren . . . 25
2.1.3 Eigenschaften der Dienste . . . 25
2.1.4 Verwaltung und Speicherung der Dienste im Zentrum . . . 25
2.2 Simulation des generischen Modells . . . 26
2.3 Optimierte Lösungsfindung . . . 26
2.3.1 Anzahl der Zentren im System . . . 27
2.3.2 Standortwahl der Zentren . . . 27
2.3.3 Zuweisung der Nutzer zu einem Zentrum . . . 27
2.3.4 Speicherung der Dienste in den Zentren . . . 27
2.3.5 Zusammenhang der Optimierungsaspekte . . . 28
2.3.6 Platzierung gemäß dem k-center und k-median Problem . . . 29
2.3.7 Gruppierung gemäß dem k-cluster Problem . . . 30
2.4 Voraussetzungen und Annahmen . . . 32

3 Anforderungsanalyse **33**
3.1 Nicht-Funktionale Anforderungen . . . 33
3.2 Funktionale Anforderungen . . . 33
3.3 Systemübergreifende Anforderungen . . . 34
3.4 Bewertungskriterien . . . 35

4 Stand der Wissenschaft **37**
4.1 Logistik - Facility Location Problem . . . 37

4.2 Logistik - Nabe-Speiche-Systeme . . . 40
4.3 Rechnernetze - Content Delivery Networks . . . 41
4.4 Ganzheitliche Ansätze . . . 45
4.5 Modellierung von Nutzerprofilen und Analyse des Speicherverhaltens in Zentren . . . 47
4.6 Zusammengefasste Gegenüberstellung bisheriger Arbeiten . . . 50
4.7 Limitierungen derzeitiger Ansätze . . . 53
4.8 Referenzalgorithmen zur Lösungsfindung in Teilbereichen . . . 53
4.8.1 Monte Carlo . . . 54
4.8.2 Zwei-Approx . . . 54
4.8.3 Greedy . . . 55
4.8.4 GRASP . . . 56
4.8.5 Partitionierende Clusterverfahren . . . 57
4.8.6 Integer Linear Programming . . . 59
4.8.7 Frameworks für Algorithmen . . . 60

5 Generisches Modell von zweistufigen Transportketten 63
5.1 Modellierung . . . 63
5.1.1 Abstraktionskonzept der Transportkette . . . 64
5.1.2 Modellierung des Nutzers . . . 67
5.1.3 Modellierung des Dienstes . . . 68
5.1.4 Modellierung des Interessen-Profils . . . 69
5.1.5 Modellierung des Zentrums . . . 72
5.1.6 Modellierung der Infrastruktur . . . 73
5.1.7 Abstrahiertes System . . . 75
5.1.8 Einbeziehung von Nebenbedingungen . . . 76
5.1.9 Zielfunktion . . . 78
5.1.10 Modell-Erzeugung . . . 82
5.1.11 Optimierung . . . 82
5.1.12 Simulation . . . 86
5.2 Spezifischer Anwendungsbereich für das Modell . . . 92
5.2.1 Spezifisches Modell für Logistik . . . 92
5.2.2 Spezifisches Modell für Rechnernetze . . . 95
5.2.3 Spezifisches Modell für Geolokalisation . . . 99
5.3 Optimierungsalgorithmus Dragoon . . . 115
5.3.1 Analyse etablierter Strategien . . . 115
5.3.2 Initialisierung . . . 118
5.3.3 Optimierung . . . 120
5.3.4 Adaption an Nebenbedingungen . . . 121
5.3.5 Spezialfall Backtrack . . . 128

6 Tragfähigkeitsnachweis und Evaluation 131
6.1 Implementierung . . . 131
6.2 Testdatensatz der Standorte und Infrastrukturen . . . 134
6.3 Testdatensatz der Profile von Nutzern . . . 136
6.4 Optimierungs- und Simulationsumgebung . . . 137

6.5 Evaluierung der Platzierungsleistung . . . 138
6.5.1 Allgemeiner Vergleich der Algorithmen bei unterschiedlichen Anwendungsgebieten . . . 138
6.5.2 Leistungsvergleich von Dragoon mit dem theoretischen Optimum . . 143
6.5.3 Einfluss von δ-Greedy auf die Platzierungsleistung im Vergleich zu Dragoon und Backtrack . . . 144
6.5.4 Überprüfung der Robustheit anhand von Leistungsdifferenzen bei der Platzierung . . . 145
6.5.5 Vergleich der Auswirkung von Platzierungsbedingungen . . . 147
6.5.6 Vergleich der Auswirkung des Optimierungskriteriums bei der Platzierung . . . 149
6.5.7 Analyse der Berechnungszeit . . . 150
6.5.8 Allgemeiner Leistungsvergleich . . . 154
6.5.9 Analyse der notwendigen Anzahl von Zentren hinsichtlich der Entfernung . . . 155
6.5.10 Analyse zur operationellen Planung . . . 156
6.6 Evaluierung der Geolokalisation . . . 157
6.6.1 Einfluss der Entfernungsfunktion und Auswahl der Vantage Points . . 158
6.6.2 Vergleich der Genauigkeit mit anderen Diensten der IP Geolokalisation . . . 159
6.7 Evaluierung des Gesamtsystems . . . 161
6.7.1 Analyse der Speicherverwaltungsstrategien in Abhängigkeit der Größe des Speichers . . . 161
6.7.2 Analyse des Speicherverhaltens der Zentren und Einfluss der Nutzerprofile . . . 163
6.7.3 Leistungspotential zwischen optimierter Platzierung und optimierter Zuweisung . . . 165
6.7.4 Pareto-Front . . . 167

7 Zusammenfassung und Ausblick 171
7.1 Beantwortung der wissenschaftlichen Fragestellungen . . . 172
7.2 Bewertung des generischen Modells und des Optimierungsansatzes . . . 173
7.3 Zukünftige Forschungsbereiche . . . 174

Literaturverzeichnis 177

Abbildungsverzeichnis 203

Tabellenverzeichnis 209

Abkürzungs- und Symbolverzeichnis 211

A Beispiel zur Motivation einer strategischen Planung von Zentren 215

B Beweis der NP Vollständigkeit des Problems 219
B.1 Angelehnte Problemstellungen im Überblick . . . 221
B.2 *k*-center . . . 221
B.2.1 Minimum *k*-center ohne Gewichte . . . 221

B.2.2 Minimum k-center mit Gewichten 221
B.3 k-cluster 222
B.3.1 Minimum k-clustering mit Gewichten 222
B.3.2 Minimum k-clustering hinsichtlich Summe bzw. Durchschnitt 222

C Beweis über die garantierte Lösungsqualität 223

D Szenarien zur Evaluation 225
D.1 Synthetisch 225
D.2 Logistik 228
D.3 Rechnernetze 230

E Evaluation 233
E.1 Vergleich der Auswirkung von Platzierungsbedingungen von Lloyd 233
E.2 Vergleich der Auswirkung von Platzierungsbedingungen von MacQueen . . . 234
E.3 Differenzen des allgemeinen Leistungsvergleichs 235

F Geolokalisation 237
F.1 Berechnung der Schnittpunkte von zwei Kreisen 237
F.2 Vergleich der Abbildungsfunktionen 238
F.3 Datensatz zur Geolokalisation 239

G Vorveröffentlichungen 241
G.1 Vorveröffentlichungen mit Bezug zur Dissertation 241
G.2 Übersicht der Vorveröffentlichungen zu Teilaspekten der Dissertation 243
G.3 Veröffentlichungen während der Arbeit an der Dissertation ohne direkten Bezug 244

Erklärung 247

1 Einleitung

Standortentscheidungen sind ein wesentliches Element der strategischen Unternehmensplanung. Ziel dabei ist den Anforderungen der Nutzer zu entsprechen und sich somit im Wettbewerb gegen andere Unternehmen durchzusetzen. Betriebsinterne Bedingungen als auch externe Anforderungen dürfen dabei nicht außer Acht gelassen werden und sind miteinander zu vereinen. Dafür wägen Entscheidungsträger Vor- und Nachteile gegeneinander ab und treffen entsprechend ihres eigenen Ermessens Entscheidungen. Diese bergen Risiken in sich und haben verschiedene Konsequenzen. Daher werden bei einer hohen Führungsverantwortung Systeme zur Entscheidungsunterstützung eingesetzt, um auf Grundlage von technischen Analysen den Einfluss von subjektiven Aspekten zu vermindern.

Die Geschichte der Standortplanung reicht dabei weit zurück. Die Abbildung 1.1 zeigt eine Methode, welche zur Standortplanung von der dänischen Hauptstadt Kopenhagen durch Bischof Absalon im Jahr 1167 eingesetzt wurde.

Abbildung 1.1: Nutzung einer stochastischen Methode für die Standortbestimmung durch Bischof Absalon vor über 850 Jahren zur Gründung der Stadt Kopenhagen [1, 2]

Hierbei wurden Nebenbedingungen vernachlässigt, wie z. B. vorhandene Infrastrukturen oder Katastrophen gefährdete Gebiete. Um möglichen anschließenden Problemen vorab entgegenzuwirken, wurden systematischere Ansätze mittels Modellen entwickelt.

Bei der betrieblichen Standortplanung mit quantitativen Aspekten ist zunächst zwischen zwei Kategorien von Organisationsformen zu unterscheiden [3]. Beim Ansatz der Standort-Zuweisung werden die Nutzer den definierten Standorten von vornherein fest zugeteilt. Demgegenüber erfolgt bei Systemen der Standort-Wahl die Entscheidung selbständig durch die Nutzer. Daher sind diese Kategorien durch unterschiedliche Modellierungsansätze zu behandeln. In der vorliegenden Arbeit wird von einem rational handelnden Nutzer ausgegangen. Dieser trifft seine Entscheidung dynamisch anhand der für ihn wichtigen Kriterien und dem größten Nutzen [4]. Bei der Modellierung des Sachverhaltes wird dies vom Entscheidungsunterstützungssystem ermittelt und die Nutzer folgen dem Vorschlag. Aus der Sicht des Betreibers hängt die Entscheidung eines Nutzers von den zur Verfügung stehenden Einrichtungen und deren Standort ab. Dadurch beeinflusst andererseits das Verhalten der Kunden die betriebliche Standortplanung [5]. Somit liegt der Fokus der Arbeit auf dem Ansatz gemäß der Standort-Wahl, wobei an die Nutzer eine entsprechende Erwartung besteht. Darüber hinaus werden beide Prinzipien berücksichtigt, um eine umfassende Planung vorzunehmen.

Die zu planenden Standorte dienen der Erfüllung von Nutzeranforderungen. Hierbei werden Transportketten zum Transfer eingesetzt. Diese stellen eine organisatorische Verbindung des Transportprozesses vom Anbieter über Zwischenstationen bis hin zu den Nutzern dar. Hierbei wird eine Steigerung der Effizienz beim Transport erreicht, indem Lieferungen auf Teilstrecken zusammengefasst werden. Durch die Verbindung mehrerer Transportketten miteinander entsteht ein ganzes Transportnetz, womit sich größere Gebiete und viele Nutzer bedienen lassen. Die Zwischenstationen in einer Transportkette übernehmen in Abhängigkeit vom Anwendungsgebiet verschiedene Aufgaben. Neben der Bereitstellung und Verteilung von den zu transportierenden Objekten dienen solche Verteilzentren zur Sicherstellung der Qualität des Dienstes und verbessern die Leistungsfähigkeit des Transportsystems durch deren Funktion als Zwischenspeicher.

Es gibt mehrere Anwendungsgebiete, in denen Transportketten und -netze Einsatz finden. Die ersten Ketten entstanden im Bereich der Logistik zur Verteilung von Waren [6]. Dabei werden große Warenmengen vom Produzent zu einem Zwischenlager befördert, von dem aus eine bedarfsgerechte Anzahl weiter an den nächsten Abnehmer geliefert wird. Die Abbildung 1.2 zeigt für den Bereich des Güterverkehrs in Deutschland die Transportleistung nach Verkehrszweigen. Bei der Betrachtung des Transportvolumens wird ersichtlich, dass eine Reduktion der Transportstrecke von nur 1 % ein erhebliches Einsparpotential an Kosten und Zeit bietet. Bei der Planung werden zudem verschiedene Standortfaktoren berücksichtigt wie z. B. Arbeitskräfte, gesetzliche Rahmenbedingungen und Steuern. Eine Befragung von Führungskräften internationaler Unternehmen zu besonders wichtigen Standortfaktoren ergab, dass über 25 % die Infrastruktur als maßgeblich für Investitionsentscheidungen erachten [7].

Das Prinzip der Transportkette kommt darüber hinaus auch bei Rechnernetzen in abstrakterer Form zum Tragen. Im Kontext des stückweisen Transfers werden digitale Daten von der Quelle über mehrere Zwischenspeicher bis hin zum Nutzer übertragen. Das Ziel besteht dabei in einem besseren Zugriff auf die Daten. Die Prognose von der Firma Cisco Systems in der Abbildung 1.3 zeigt die voraussichtliche exponentielle Steigerung des Datenaufkommens durch Privatanwender in den nächsten Jahren.

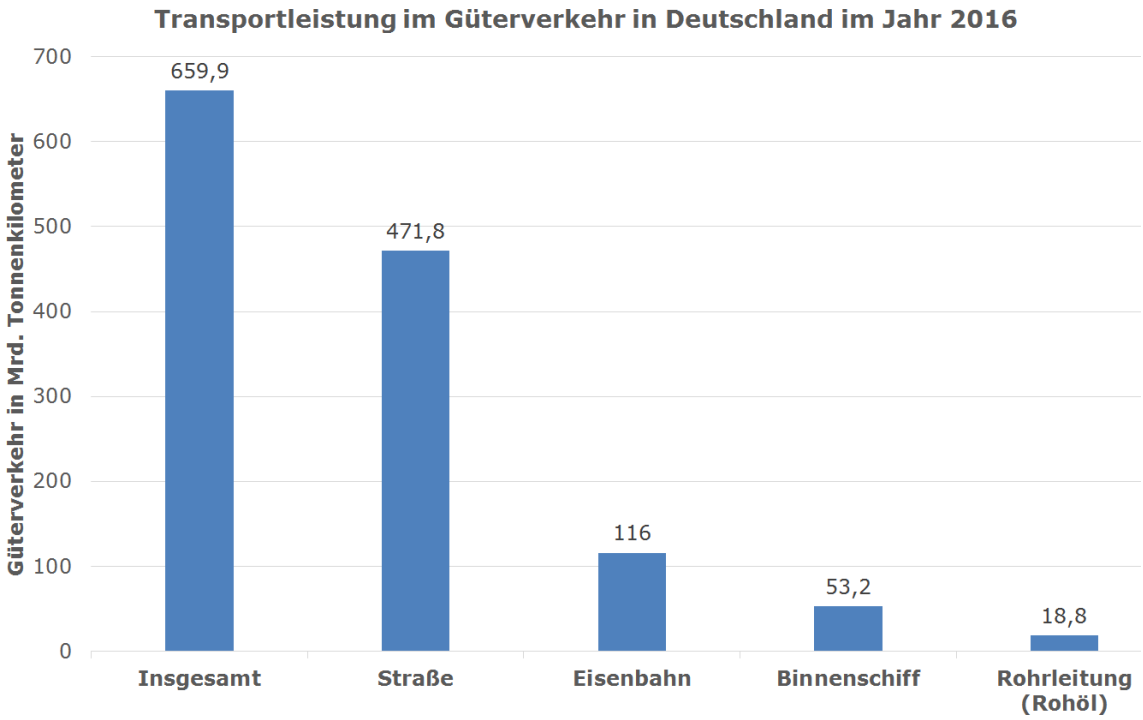


Abbildung 1.2: Transportleistung in Deutschland nach Verkehrszweigen [8]

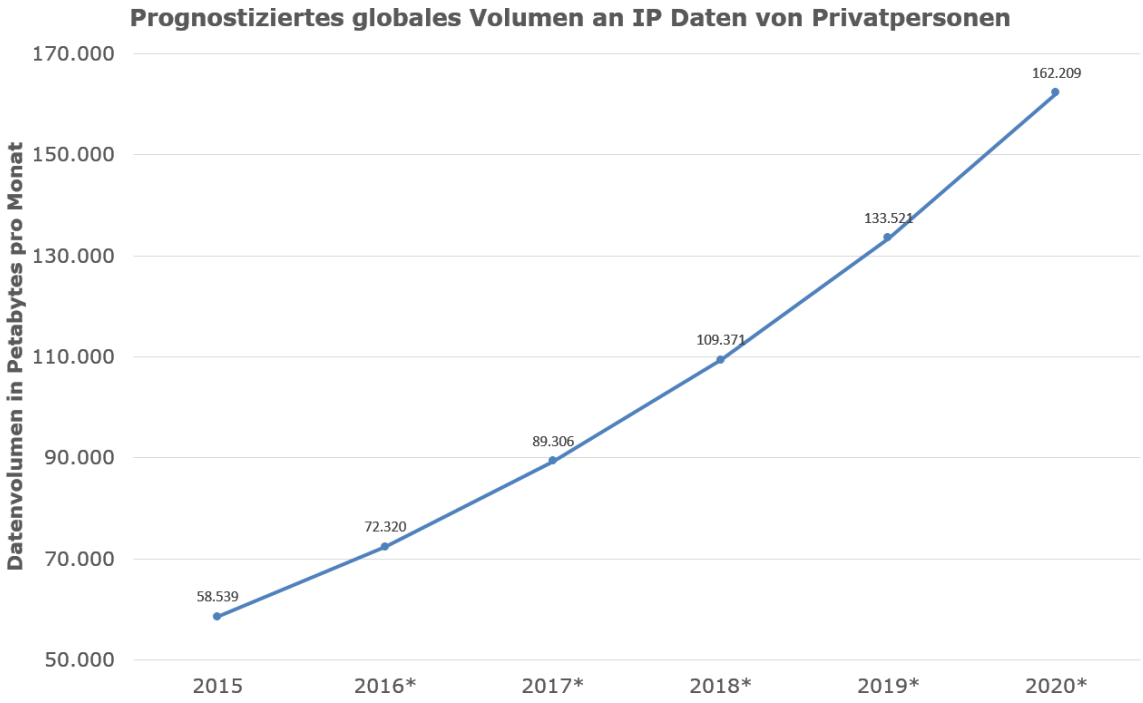


Abbildung 1.3: Prognostizierter Anstieg des globalen Datenverkehrs von Privatpersonen [9]

Um diese Datenmengen zu bewältigen, sind effiziente Übertragungskonzepte und Gesamtsystemlösungen notwendig.

In Anbetracht der Planung und Optimierung von solchen Transportketten ergeben sich komplexe Problemstellungen. Die Bestimmung angemessener Standorte für die Zwischenstationen stellt dabei eine wesentliche Herausforderung für das Management dar. Dabei sind speziell die Zuordnungen der verschiedenen Transferleistungen untereinander zu beachten. Dies hat einen direkten Einfluss auf die Leistungsmerkmale wie Dienstqualität, zukünftige Kosten und Konkurrenzfähigkeit. Für ein effizientes System sind Nebenbedingungen und zusammenhängende Einflussfaktoren zu berücksichtigen, sodass eine Integration in das Umfeld erreicht wird. In der Literatur werden mehrere Lösungsansätze beschrieben, die für einzelne Teilprobleme verschiedene Eigenschaften während der Optimierung betrachten und Teilaspekte genauer analysieren. Jedoch fehlt es an einem allgemeinen Ansatz für verschiedene Anwendungsgebiete, um Erkenntnisse in einem Bereich direkt in einen anderen Szenario zu nutzen. Zudem ist für eine umfassende Lösung eine übergreifende Betrachtung zusammenhängender Eigenschaften und Beziehungen vorzunehmen. Dies impliziert eine gleichzeitige Optimierung hinsichtlich mehrerer Zielkriterien, wobei der Einfluss von vor- sowie nachgelagerten Prozessen berücksichtigt wird. Darüber hinaus ist insbesondere der Einfluss des Nutzerverhaltens mit in die Planung des Transportsystem ein zu beziehen. Die Modellierung des Nutzerverhaltens und die flexible Einbindung in bestehende als auch neue Modelle gehört zu den gegenwärtigen Forschungsschwerpunkten in diesem Bereich. Weiterhin mangelt es für die Platzierung von zentralen Knotenpunkten anhand einer vorgegebenen Infrastruktur und Graphen an effizienten Lösungsalgorithmen. Die Problemstellungen gleichen sich in vielen Anwendungsbereichen auf Grund des ähnlichen Funktionsprinzips in Bezug auf die Verteilzentren. Daher lassen sich grundlegende Lösungsansätze für mehrere Bereiche adaptieren.

In der vorliegenden Dissertation wird für die zuvor beschriebene Problemstellung ein Modell erstellt. Dieses ist durch Parametrierung flexibel anpassbar und lässt sich für mehrere Anwendungsgebiete einsetzen. Es werden realistische Nebenbedingungen auf die Problemstellung betrachtet, um einen höheren Detailgrad für die praktische Anwendbarkeit zu erreichen. Dies beinhaltet verschiedene Eigenschaften und Prioritäten von den einzelnen Bestandteilen und Verbindungen im Transportsystem. Für eine optimierte Planung ergibt sich unter anderem das Ziel der Minimierung von Transportweg und Transportzeit, sodass die Nutzer eine geeignete Anbindung und komfortablen Service erhalten. Aus Sicht des Betreibers ist ein optimiertes Kostenverhältnis zwischen den lokalen Betriebskosten der zentralen Standorte und den Transportkosten zur Versorgung zu erreichen. Dies geht einher mit der Anzahl von zentralen Knotenpunkten und deren Verwaltung zur Erreichung einer definierten Dienstqualität. Dadurch ergibt sich ein multikriterielles Optimierungsproblem. Zudem erfolgt eine Betrachtung weiterer Einflussfaktoren im Vergleich zu bestehender Literatur wie dem Verhaltensprofil von Nutzern. Diese werden zur weiteren Steigerung der Leistungsfähigkeit und der Dienstqualität in die Modellierung und Optimierung des Systems mit einbezogen.

Der vorliegende Beitrag bildet das Verhalten über dynamische Nutzerprofile nach. In den betrachteten Szenarien stellen mehrere Nutzer individuelle Anfragen an die Zentren in

einem Transportsystem. Ein Nutzer steht dabei vor dem diskreten Entscheidungsproblem, welches Zentrum des Systems zu berücksichtigen ist und entsprechende Anfragen bearbeitet. Hintergrund ist eine zielgerichtete und den Anforderungen der Nutzer entsprechende Belieferung bei gleichzeitiger Vereinfachung der Abarbeitung von Anfragen an den zentralen Knotenpunkten. Dadurch werden nutzbringende Auswirkungen auf verschiedene Laufzeit- und Langzeiteffekte im Gesamtsystem erreicht. Insbesondere lassen sich dabei durch das Modell stochastische Einflüsse analysieren. Die Intention einer optimierten Lösung besteht somit in der Erreichung einer hohen Qualität für die Nutzer bei vermindertem Einsatz von Ressourcen für den Anbieter. Weiterhin wird durch die umfassendere Betrachtung eine zusammenhängende Optimierung vorgenommen und somit ein Bezug zu angrenzenden Rahmenbedingungen für das betrachtete Problem hergestellt. Folglich wird mit dem Ansatz eine wissenschaftliche Grundlage geschaffen, um über verschiedene Analysen adäquate Entscheidungen zu treffen.

Die Evaluation erfolgt mittels empirischer Auswertung der Optimierungsergebnisse und simulierter Experimente von verschiedenen realitätsnahen Szenarien. In einer Serie von Tests werden einzelne Aspekte genauer analysiert und Rückschlüsse für das Gesamtkonzept gezogen. Die gewonnenen Erkenntnisse aus den detaillierten Analysen dienen der Entwicklung eigener Optimierungsverfahren, die mit verschiedenen, existierenden Algorithmen verglichen werden. Die Optimierung erfolgt in Abhängigkeit eines oder mehrerer Zielkriterien, die je nach Anwendungsgebiet unterschiedlich ausfallen. Typische Kriterien sind die Minimierung der maximalen Transportzeit, durchschnittliche Entfernung und Leistungsmaximierung sowie Kostenminimierung. Zudem besteht das Interesse an der Zuweisung von ähnlichen Bedarfsforderungen der Nutzer an einem Zentrum unter Berücksichtigung deren Profile. Darüber hinaus lässt sich durch vergangene Anfragen der Einfluss der Historie messen. Der Leistungsvorteil wird anhand der Erfolgsquote von Zwischenspeicher-Verwaltungsstrategien nachgewiesen. Hierbei erfolgt auch eine Evaluation der verschiedenen Strategien in Zusammenhang mit dem Anwendungsgebiet.

Das Ergebnis der Dissertation ist ein generisches Modell zur Lösung der Problemstellung mit flexibler Anpassbarkeit. Der grundlegende Sachverhalt der Planung basiert auf den k-center und k-cluster Problemen. Hierbei sind k zentralisierte Standorte zu bestimmen, um eine vorgegebene Menge an weit verteilten Knotenpunkten bestmöglich zu erreichen. Neben Lösungsvorschlägen werden angepasste Optimierungsalgorithmen entwickelt und beschrieben. Durch die Einbeziehung der Verhaltensprofile in die optimierte Standortplanung wird eine Leistungssteigerung erzielt. Darüber hinaus erfolgen Analysen zur Problemstellung, die verschiedene Charakteristika aufzeigen. Die erhaltenen Erkenntnisse unterstützen das Management in einer effizienten Entscheidungsfindung. Das generische Modell und die entwickelten Optimierungsalgorithmen sind neben den genannten Bereichen der Logistik und Rechnernetzen in weiteren Anwendungsgebieten nutzbar. Dazu gehören Datenanalysen mittels Data Mining, Funkzellenbestimmung im Mobilfunk zur Bereichsabdeckung und Planung von Lieferstationen für Drohnen. Zudem lassen sich im Rahmen der Geolokalisation Messpunkte zur Standortbestimmung von Netzkomponenten ermitteln.

1.1 Einordnung der Arbeit

Im Zuge der immer stärkeren Vernetzung verschiedener Bereiche kommt es bei Transportketten mit der Thematik Industrie 4.0 zu der Entwicklung sogenannter *Connected Supply Chain*. Dies ermöglicht die Optimierung der Prozesse über die gesamte industrielle Wertschöpfungskette hinweg. Die Verbindung zwischen digitaler und physischer Welt wird auch als Cyber-physisches System bezeichnet und ermöglicht damit neuartige Funktionen und Dienstleistungen. Diese Systeme bieten die Fähigkeit an verschiedenen Punkten des Transportnetzes Informationen über den Status der Dienste zu erhalten. Durch die bessere Nachvollziehbarkeit lassen sich Optimierungsmaßnahmen einleiten, um das Transportsystem effizienter und reaktionsfähiger zu gestalten.

Die vorliegende Arbeit ist spezifisch für die Planung des Transportsystems und integriert sich durch die umfassende Betrachtung in den Gesamtkontext der *Connected Supply Chain*. Zum Transportsystem gehört die Planung von Zwischenstationen mit entsprechender Zuweisung verantwortlicher Bereiche zur Belieferung. Als essentieller Bestandteil logistischer Systeme liegt der Fokus dieser Arbeit auf dem Transportsystem mit einer detaillierten Lösung. Zur Einordnung der Arbeit stellt die Abbildung 1.4 beeinflussende Planungsebenen dar, welche bei der Bearbeitung des Themas zu beachten sind, um eine integrierte Lösung zu entwickeln. Die Beziehungen zwischen den angrenzenden Planungsebenen untereinander werden auf Grund der Übersichtlichkeit vernachlässigt.

Beginnend mit der Identifizierung der Nutzer werden in diesem Planungsbereich die spezifischen Anforderungen zur Belieferung durch die Transportkette ermittelt. Mit deren Standorten als Bedarfsträger haben sie einen maßgeblichen Einfluss auf die Planung des Systems. Die Ebene der Infrastruktur übernimmt die Verbindung der verschiedenen Standorte von Nutzern, Produktion und Zwischenstationen. In Abhängigkeit der verfügbaren Transferwege wird die Planung des Transportsystems vorgenommen und die Infrastruktur gegebenenfalls um zusätzliche Verbindungen erweitert. Der Bereich der Produktionsplanung stellt mit den zugehörigen Standorten die Ausgangspunkte einer Transportkette dar. Daher beeinflusst die Identifizierung möglicher Produktionsstandorte die Planung von Zwischenstationen, zumal diese durch die Zwischenspeicherung selbst die Fähigkeiten von Zwischenstation besitzen. Das Teilgebiet operative Wegewahl verwendet unter anderem aktuelle Verkehrsinformationen, um bei der regelmäßigen Belieferung der Bedarfsträger eine zeitlich und kosteneffiziente Route zu identifizieren. Solche Verkehrsinformationen lassen sich auch vorab als zusätzliches Wissen bei der Planung des Transportsystems mit einbeziehen. Dadurch lassen sich vermeidbare Schwierigkeiten von vornherein umgehen. Im Zusammenhang mit dem Transportsystem und der Wegewahl steht auch der Bereich der Kapazitätsplanung. Hierbei wird sicher gestellt, dass die angeforderten Mengen mittels dem Transportsystem zu den entsprechenden Standorten gelangen. An den Zwischenstationen ist dafür ausreichend Kapazität zur Verfügung zu stellen, sodass gegebenenfalls auch zusätzliche Reserven vorhanden sind.

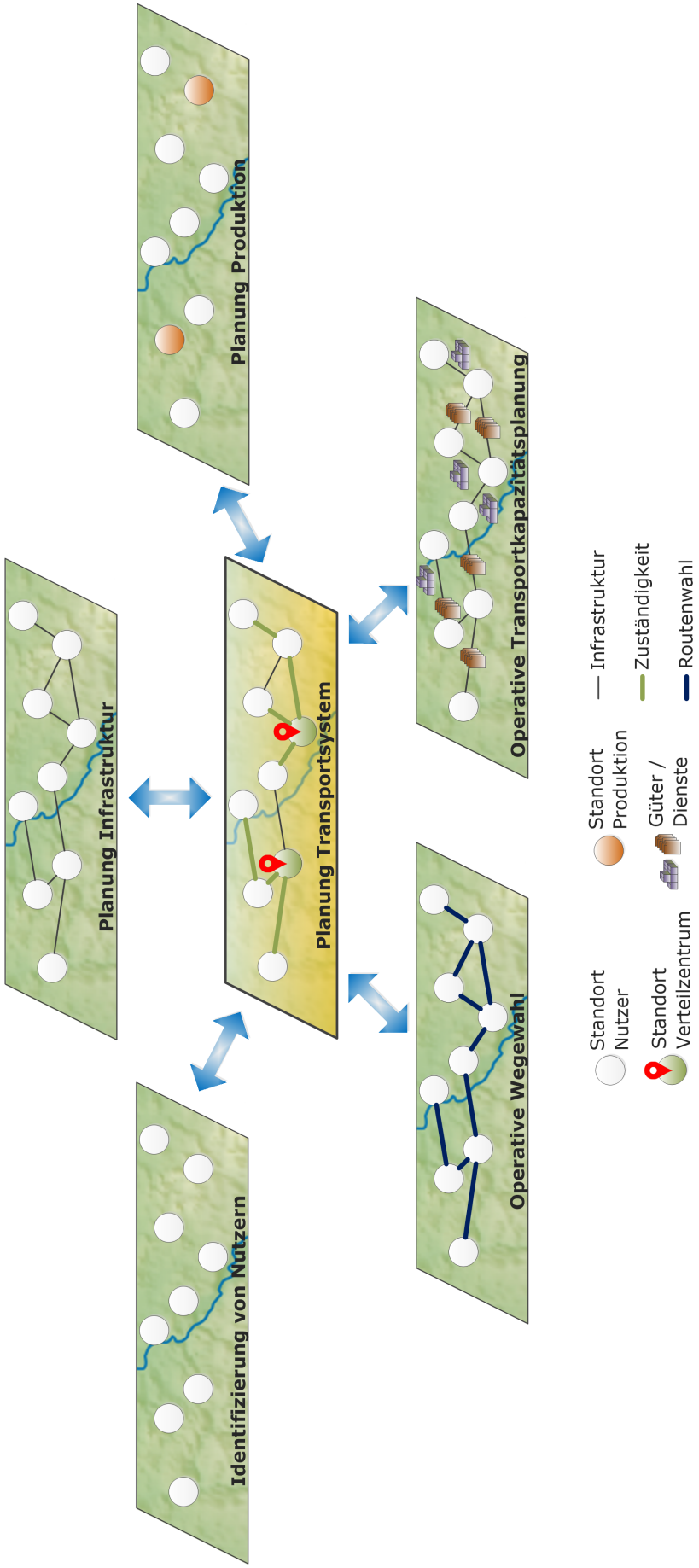


Abbildung 1.4: Einordnung dieser Arbeit im Zusammenhang mit weiteren Planungsbereichen bei Transportketten

1.2 Anwendungsgebiet und Motivation

Dieser Abschnitt beschreibt die in der vorliegenden Arbeit betrachteten Anwendungsgebiete und geht auf verschiedene Nebenbedingungen ein. Dabei werden grundlegende Begriffe erläutert, deren Bedeutung für das Verständnis der Arbeit relevant sind. Zunächst erfolgt eine allgemeine Betrachtung des Themengebietes Transportkette. Die aufgezeigten Eigenschaften werden anschließend in zwei Anwendungsgebieten dargelegt. Hierbei erfolgt eine detaillierte Beschreibung von jeweils einem Beispiel in den unterschiedlichen Bereichen Logistik und Rechnernetze. Abschließend wird das Ziel von der Arbeit erläutert und es erfolgt eine vereinheitlichte Definition von Begrifflichkeiten.

1.2.1 Transportkette und -system

Eine Transportkette [10] stellt in Anlehnung an DIN 30781 [11, 12] eine Folge technisch und organisatorisch verknüpfter Vorgänge dar, bei denen Güter von einer Quelle zu einer Senke befördert werden. Dabei erfolgt der Transport über zentrale Zwischenstationen zur Bündelung oder Auflösung des Güterflusses. Die zu transportierenden Gegenstände und Produkte werden als Verkehrsobjekte bezeichnet. Der Begriff Kette bezieht sich auf die Folge von Transport- und Aufenthaltsvorgängen von Verkehrsobjekten zwischen dem Start- und dem Zielort über Verteilzentren als Zwischenstation. Die zentralen Verteilstationen bedienen die Anforderungen des jeweils zugehörigen Einzugsgebietes beziehungsweise das entsprechende regionale Nahgebiet [13]. Das Ziel der Nutzung von Verteilzentren besteht in der Optimierung der notwendigen Ressourcen des Transportsystems bei gleichzeitiger Erfüllung hoher Qualitätsanforderungen. Darüber hinaus besteht die Motivation in der Reduktion der Transportzeit und -kosten. Allgemein ist es dafür von Vorteil, wenn die Verteilzentren der Transportkette jeweils nahe bei den Senken liegen. Sofern sich die Zentren bei Bedarfsanfragen näher am Zielort befinden als die Quelle selbst, wird durch die Zwischenlagerung die verbleibende Transportstrecke reduziert. Weiterhin ermöglicht der Einsatz von Verteilzentren eine quantitative Speicherung der Güter, sodass ein Ausgleich der Bedarfsschwankungen mehrerer Senken erfolgen kann. Indem der Bedarf von mehreren Verteilzentren gedeckt wird, findet ein Lastausgleich und eine Kapazitätssteigerung statt. Zusammenfassend lassen sich durch den Einsatz von Transportketten mehr Senken auf eine effiziente Weise erreichen.

Transportketten lassen sich auf Grundlage der Umschlaghäufigkeit an den Zwischenstationen in ein- und mehrstufige Systeme unterscheiden. Bei einstufigen Transportketten erfolgt kein Umschlag der Güter in einem Verteilzentrum. Es besteht somit eine direkte Verbindung zwischen Quelle und Senke. Eine mehrstufige Transportkette liegt vor, wenn die Güter auf indirektem Weg über Zwischenstationen zum Zielort gelangen. Die kürzeste Kette mit Verteilzentren ist somit die zweistufige Transportkette. Der Schwerpunkt dieser Arbeit liegt auf der Planung und Optimierung der Verteilzentren zur Zwischenspeicherung. Daher erfolgt eine Fokussierung auf zweistufige Transportketten, welche sich von der Quelle über ein Verteilzentrum bis hin zu den Senken erstrecken. Darüber hinaus wird die Erweiterung zu mehrstufigen Ketten in Betracht gezogen.

Bei der Planung von Transportsystemen ergeben sich in den verschiedenen Anwendungsgebieten komplexe Fragestellungen. Mit der Betrachtung mehrerer Ketten im Zusammenhang entsteht durch das gemeinsame und kooperative Zusammenwirken der Betreiber ein Transportnetz. Dieses bildet unter Beachtung von äußeren Einflussfaktoren ein umfangreiches System. Dazu werden in der vorliegenden Arbeit verschiedene Probleme aufgezeigt, analysiert und gelöst. Unter anderem besteht bei der Planung das *Facility Location Problem* (FLP). Hierbei sind aus einer Menge von verschiedenen Standortmöglichkeiten für Verteilzentren mehrere Standorte auszuwählen. Die in diesem Zusammenhang vom Supply-Chain-Management zu treffenden Entscheidungen zur Umsetzung sind kostenintensiv und haben weitreichende Folgen [14]. Die konzipierten Systeme sind über lange Zeiträume im Einsatz, sodass Beschlüsse nachhaltige Langzeiteffekte aufweisen. Weiterhin haben die Festlegungen von der Planung Auswirkung auf das anschließende operative Systemverhalten.

Im Folgenden wird die Nutzung von Verteilzentren bei Transportketten in den beiden unterschiedlichen Anwendungsgebieten Logistik und Rechnernetze motiviert. Anschließend erfolgt eine detailliertere Betrachtung eines Beispiels aus dem jeweiligen Bereich. Dabei werden durch die gezielte Beschreibung Gemeinsamkeiten und Unterschiede herausgestellt. Die Auswahl der beiden Gebiete erfolgt auf Grund der einfachen Verständlichkeit, der hohen Divergenz in der Ausprägung der Eigenschaften und der unterschiedlichen Sichtweise auf eine Transportkette. Bei vergleichender Betrachtung der Beispiele zeigen sich im Detail diese signifikanten Gegensätze. Werden die Prozesse jedoch abstrahiert, ergeben sich Gemeinsamkeiten in der Art der Eigenschaften, welche lediglich eine unterschiedliche Ausprägung in den Domänen besitzen.

1.2.2 Motivation zum Einsatz von Verteilzentren

Im Bereich der Logistik dienen Verteilzentren einer Transportkette der Belieferung einer vorgegebenen Menge an Kunden. Hierbei besteht das Ziel in der Bündelung von Transportstrecken, sodass sich eine Effizienzsteigerung ergibt und in der Gesamtheit eine Reduktion der Transportstrecke und -zeit resultiert. Abbildung 1.5 verdeutlicht dazu die Erweiterung der einstufigen Transportkette zu einer zweistufigen Kette.

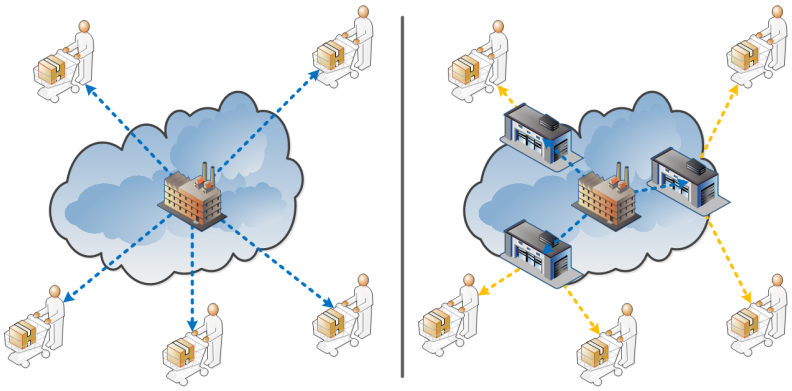


Abbildung 1.5: Erweiterung von einer einstufigen zu einer zweistufigen Transportkette im Bereich Logistik

Es werden große Mengen von Gütern des Produktionsstandortes zu einem Warenlager transportiert. Dieses gibt in Abhängigkeit des Bedarfs kleinere Mengen mit höherer Häufigkeit an mehrere Konsumenten ab. Wenn in einem bestimmten Zeitraum ein Konsument weniger Bedarf aufweist und dafür ein anderer einen höheren, dann wird durch die Zwischenlagerung ein Mengenausgleich erreicht. Somit werden kurzfristige Differenzen im Bedarf gemildert und Fehlmengen durch die vorgehaltene Kapazität im Lager ausgeglichen. Dies führt zu einer Verbesserung der Planungsfähigkeit und der Dienstqualität.

Im Bereich der Rechnernetze lassen sich *Content Delivery Network* (CDN) als eine Transportkette für Daten auffassen. Ein CDN ist ein Rechnernetz von weit verteilten Servern, welche über das Internet miteinander verbunden sind. Es bildet eine Overlay-Struktur zur Speicherung und Verteilung von digitalen Inhalten. Basierend auf einem physischen Netz wird dabei eine virtuelle Infrastruktur erstellt. Beim digitalen Transfer kopieren Server Daten von einer Quelle und stellen diese bei Bedarf zur Verfügung. Der Prozess des Kopierens wird auch als Spiegeln bezeichnet, wodurch die Server als Spiegel-Server benannt sind. Die Vervielfältigung der Daten verbessert nicht nur die Verfügbarkeit, sondern steigert auch die Leistungsfähigkeit bei der Beantwortung von Nutzeranfragen [15, 16]. Die Abbildung 1.6 zeigt dazu die Erweiterung des Systems um eine Zwischenstufe bei der Datenübertragung.

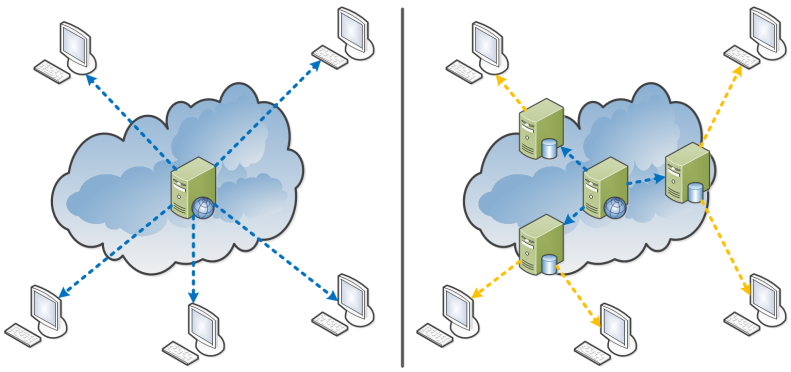


Abbildung 1.6: Erweiterung von einer einstufigen zu einer zweistufigen Transportkette im Bereich Rechnernetze

Neben der Lastverteilung auf mehrere Spiegel-Server, wird eine Steigerung der Transferrate erreicht. Durch eine Zuweisung der Nutzer zu den nächstgelegenen Servern besteht eine kürzere Entfernungen zwischen den Kommunikationspartnern. Mit der Verringerung der Verbindungslänge ergibt sich auch eine geringere Zugriffszeit. Weiterhin führt die Verteilung und Zwischenspeicherung der Daten zu einer reduzierten Netzlast im Gesamtsystem bei wiederholter Abfrage.

1.2.3 Anwendungsgebiet der Logistik

Im Folgenden dient eine Firma mit Baumärkten und einer zweistufigen Transportkette als Vorlage für ein FLP im Bereich Logistik. Zur Erläuterung der Problemstellung wird

das folgende, hypothetische Szenario betrachtet. Das Management der Baumarktkette beschließt die Erweiterung seiner Geschäfte zur Gewinnsteigerung. Hierzu expandiert die Firma in weitere Länder und exportiert verschiedene Produkte aus Holz. Die Herstellung der Waren erfolgt in Fabriken in Übersee. Von dort aus werden die produzierten Güter in großen Mengen zu zentralen Zwischenlagern in der Nähe von lokalen Baumarkt Filialen transportiert und gelagert. Anschließend erfolgt die anforderungsgerechte Aufteilung und Belieferung dieser entsprechend deren Bedarf. Den zentralen Zwischenlagern sind hierbei mehrere Filialen zugeordnet.

In dem betrachteten Beispiel der Logistik werden Verkehrsobjekte in Form von Stückgütern transportiert. Diese sind vorab in den Zwischenlagern aufzubewahren, sodass bei einer Anfrage die angeforderten Mengen sofort an die Bedarfsträger abgegeben werden können. Die zu transportierenden Mengen von den Zwischenlagern zu den Filialen sind dabei wesentlich kleiner als die Mengen von der Produktion zu den Zwischenlagern. Der Warentransport zu den Filialen erfolgt dadurch auch häufiger als zum Zwischenlager. Somit lassen sich längere Strecken zwischen dem Produktionsstandort und dem Zwischenlager hinnehmen zugunsten kurzer Verbindungen vom Zwischenlager zu den Filialen. Zudem wird der Teil der Transportstrecke zwischen Produktion und Zwischenlager für eine Belieferung der Filialen zusammengefasst. Die Zwischenlager dienen somit zur Reduktion der Transportwege und Lieferzeit zu den Filialen. Durch die quantitativ Zwischenlagerung werden die Auswirkungen von Bedarfsschwankungen abgemildert und eine kontinuierliche Versorgung gefördert. Weiterhin verbessert sich die Fähigkeit zur „Just-In-Time“ Belieferung [17, 18], einer schnellen und rechtzeitigen Erfüllung einer Bestellung.

Das Ziel besteht in einer optimierten geographischen Platzierung der Lager, sodass die Lieferwege und -zeiten der einzeln zugeteilten Filialen minimiert werden. Der Transport der Waren erfolgt über das Straßennetz, wobei diese Infrastruktur im Vergleich zu den Graphenstrukturen von Rechnernetzen stärker vernetzt ist. Diese besitzen eine dichtere Struktur mit mehr Verbindungen. Bei der Planung von Zentren besteht eine nahezu freie Standortwahl, wobei eine adequate Anbindung an die Infrastruktur mit schnellen Verbindungen wichtig ist. Kurze Strecken zur Anbindung an das Straßennetz lassen sich bei der Standortwahl zur Erschließung mit einplanen. In der Abbildung 1.7 sind auf der linken Seite zu versorgende Standorte durch grüne Punkte markiert. In diesem Beispiel sind es 143 Städte in der Schweiz, die alle mehr als 10.000 Einwohner aufweisen. Auf der rechten Seite der Abbildung sind für die Problemstellung 10 Zentren geplant und in rot markiert. Die Zuweisung der Standorte zu den Zwischenlagern wird durch die blauen Linien angezeigt.

Für die Planung sind die Standorte der Filialen sowie die geografischen und infrastrukturellen Gegebenheiten im Voraus bekannt. Auf Grund der initialen Budget-Planung des Managements lässt sich die Menge an Zwischenlagern vorab auf eine Höchstanzahl eingrenzen. In Abhängigkeit der Kosten an den verschiedenen Standorten der Zwischenlager kann es vorteilhaft sein, nur an wenigen Orten die Zwischenlager zu eröffnen und dafür höhere Transportkosten zu akzeptieren, bis hin zu vielen Zwischenlagern mit kurzen Transportwegen. Die Transportkosten ergeben sich aus der Entfernung einer Filiale zu dem entsprechend zugewiesenen Zwischenlager. Hierbei sind

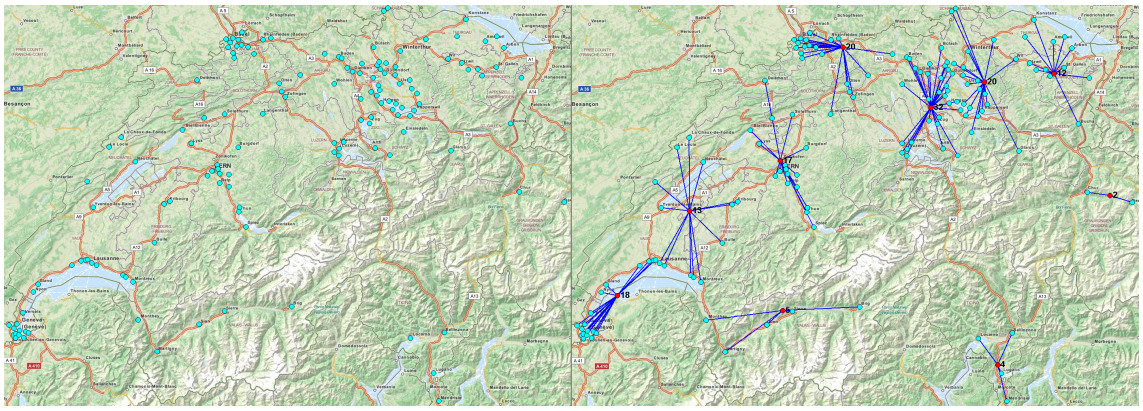


Abbildung 1.7: Szenario eines FLP anhand der 143 größten Städte in der Schweiz mit der Planung von 10 zentralen Zwischenlagern (Kartendaten © OpenStreetMap-Mitwirkende, CC-BY-SA)

die Transportkosten gegenüber den operationellen Betriebskosten der Zwischenlager abzuwägen. Weitere Vor- und Nachteile eines Standortes sind entsprechend bei der Planung zu berücksichtigen.

1.2.4 Anwendungsgebiet der Rechnernetze

Im Folgenden dient das System eines CDN als Beispiel im Bereich digitaler Transportketten. Die Beschreibung der Problemstellung erfolgt anhand des nachstehenden, hypothetischen Szenarios. Der Betreiber eines CDN beabsichtigt seine bestehende Speicher-Infrastruktur mit weiteren Servern zu vergrößern und an die aktuelle Situation anzupassen. Zur Lastverteilung von Nutzeranfragen werden die entsprechenden Daten vom ursprünglichen Anbieter auf mehrere Server des CDN Betreibers gespiegelt und die Anfragen verteilt. Dazu wird eine definierte Menge an Servern in der Netzinfrastruktur an strategischen Punkten eingerichtet.

Zu Beginn des Planungsprozesses zur optimierten Positionierung der Server ist eine Analyse der bestehenden Netz-Infrastruktur und der daran angeschlossenen Clients notwendig. Die Infrastruktur wird durch den *Internet-Service-Provider* (ISP) verwaltet und ist somit vorgegeben. Bei realen Netzen ist nicht jeder Knoten mit jedem anderen direkt verbunden, sodass eine Ende-zu-Ende-Verbindung über mehrere Punkt-zu-Punkt-Verbindungen erfolgt. Somit besteht im Bereich Rechnernetze eine Infrastruktur, welche im Vergleich zu Straßennetzen spärlicher vernetzt ist. Weiterhin sind reale Infrastrukturen von Weitverkehrsnetzen auf Grund der Vernetzung mit anderen Netzen nicht hierarchisch strukturiert. Einzelne Verbindungen haben unterschiedliche Bandbreiten, Latenzen und Auslastungen. Die Clients sind über sogenannte Gateway-Knoten an das Netz angeschlossen. Alle Anfragen der Nutzer werden über einen solchen Knoten an den Verbund von Spiegel-Servern gesendet. Dabei ist den verschiedenen Clients entsprechend ihres Service-Level-Agreements eine Priorität zugeordnet.

Die Abbildung 1.8 zeigt auf der linken Seite das Backbone Netz der Firma Dialtelecom in der Tschechischen Republik [19]. Dieses enthält 124 Knoten und 126 Verbindungen. Die geographischen Standorte der Netz-Knoten, welche einen Anschluss neuer Verbindungen

ermöglichen, sind jeweils als grüne Punkte dargestellt. Auf der rechten Seite sind 10 Standorte für die Errichtung von Servern ausgewählt. Diese bilden das System eines CDN. Die Zuweisung der Nutzer zu den jeweils nächstgelegenen Servern, in Bezug auf die Infrastruktur, wird durch die blauen Linien angezeigt.

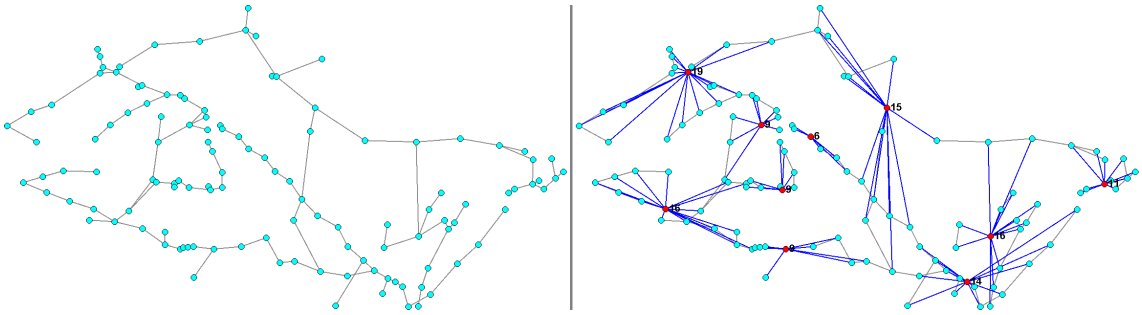


Abbildung 1.8: Beispiel eines CDN anhand des Rechnernetzes in der Tschechischen Republik mit 124 Knoten und der Selektion von 10 Standorten zur Anbindung der Server (Netz: Dialtelecom (Cz) [19])

Bei der Standortwahl sind die Server in dem Netz entsprechend so zu platzieren, dass die Nutzer eine geringe Zugriffszeit mit hoher Bandbreite auf den Server haben [20, 21]. Dazu ist die Verbindung auch hinsichtlich einer kurzen Entfernung zwischen Nutzer und Server zu optimieren. Die Platzierung eines Servers nahe dem Client bezieht sich hierbei nicht auf die geografische Nähe, sondern auf die infrastrukturelle Anbindung [22]. Nur weil zwei Standorte geografisch nah beieinander liegen, muss die Verbindungsstrecke für eine Datenkommunikation auf Grund der Infrastruktur nicht kurz sein. Daher hat die Standortbestimmung für die Server auf Grundlage der infrastrukturellen Verbindungen zu erfolgen. Dieses Vorgehen verbessert die *„Quality-of-Service“* (QoS) [23, 24] und *„Quality-of-Experience“* (QoE) [25] für die Nutzer. Die mehrfachen Datenübertragungen zwischen den Nutzern und den Spiegel-Servern erfolgen über optimierte Verbindungen mit kurzen und wenigen Verbindungswegen. Die Server fassen die Datenanfragen der Nutzer an die ursprüngliche Quelle zusammen und nutzen bei wiederholten Anfragen den eigenen Speicher. Somit führen die optimierten Verbindungspfade und die Zwischenspeicherung der Daten auf den Servern zu einer Reduzierung der Netzlast Zudem verringern sich aus Sicht des Infrastrukturbetreibers gegebenenfalls die Transitkosten durch die Speicherung der Daten im Netz des jeweiligen ISP, da weniger Daten zum eigenen Netz zu transferieren sind.

Die Funktionsweise des CDN erfolgt transparent für die Nutzer. Bei digitalen Transportketten werden Datenpakete versendet, wobei sich gespeicherte Inhalte eines Servers beliebig oft abrufen lassen. Die in dieser Arbeit betrachteten CDN gehören zur Kategorie der nutzergetriebenen CDN, sogenannten Community-Based-CDN. Hierbei speichern die Server die angeforderten Daten reaktiv. Dabei erfolgt keine Vorabspeicherung von Inhalten, sondern erst bei Anfrage durch einen Nutzer fordert der Server fehlende Daten bei einer Quelle an. Nach diesem Prinzip funktionieren Systeme wie CORAL [26] und MirrorBrain [27].

In der Planungsphase stellt der Anschluss der Spiegel-Server an die Netzinfrastruktur ein komplexes kombinatorisches Optimierungsproblem dar. Diese Aufgabe ist dahingehend

vergleichbar mit dem Beispiel aus der Logistik. Um eine geplante Dienstqualität zu erreichen, ist es notwendig, die entsprechende Anzahl von Servern für eine bestimmte Infrastruktur zu identifizieren. Ein angepasstes Modell zur Systemplanung und die Analyse notwendiger Server in der Infrastruktur, als auch die gespeicherten digitalen Inhalte auf einem Spiegel-Server, ermöglichen softwaregestützte Entscheidungen zu treffen.

1.2.5 Motivation für generischen Ansatz

Die beiden zuvor genannten Beispiele sind entsprechend aufbereitet, damit die Gemeinsamkeiten im allgemeinen Funktionsprinzip und die spezifischen Unterschiede in der jeweiligen Ausprägung ersichtlich werden. Durch die vergleichende Betrachtung der beiden Beispiele aus den Anwendungsbereichen Logistik und Rechnernetze lassen sich grundlegende Eigenschaften feststellen. Dies betrifft insbesondere das systematische Vorgehen beim Transfer der Verkehrsobjekte und die Standortermittlung der Zwischenstationen. Abweichungen ergeben sich durch die Eigenschaften der zur transferierenden Verkehrsobjekte. Im Folgenden sind für die beiden Beispiele einige Gemeinsamkeiten und Unterschiede aufgeführt:

Gemeinsamkeiten:

- Inhärentes Funktionsprinzip der Transportkette
- Bündelung von Transfers mittels Zwischenstationen
- Bestimmung angemessener Standorte für die Zwischenstationen
- Zuweisung der Bedarfsträger zu den Stationen
- Vergleichskriterium hinsichtlich einer hohen QoS

Unterschiede:

- Art der zu übertragenden Verkehrsobjekte
- Vervielfältigung von Verkehrsobjekten
- Standortmöglichkeiten der Zwischenstationen
- Ausprägung der Infrastruktur
- Vorgehensweise bei der Zwischenlagerung der Verkehrsobjekte
- Nebenbedingungen der Anwendungsgebiete

In den genannten Domänen gibt es weitere Teilbereiche, in denen diese Herausforderungen von entscheidender Bedeutung sind. An dieser Stelle wird zunächst auf einige Szenarien für FLP hingewiesen. Zum Einsatzbereich gehören Standortbestimmungen für Katastrophenhelfer, die wissen müssen, wo Versorgungslager mit überlebenswichtigen Waren nach einer Naturkatastrophe aufzustellen sind. Weiterhin hat zur Einhaltung von Reaktionszeiten eine angepasste Planung von Stationen für Rettungsdienste und

Feuerwehren zu erfolgen. Nicht zuletzt sind im Zuge der Entwicklung von Drohnen, Paket- und Ladestationen an zentralen Orten aufzustellen, um die Auslieferung von Waren zu übernehmen [28].

Neben den Bereichen Logistik und Rechnernetze bestehen zusätzliche Anwendungsgebiete. Im Rahmen der Geolokalisation sind Messpunkten zur Entfernungsermittlung für die Triangulation und Lateration unbekannter Punkte festzulegen [29]. Bei Telekommunikationsunternehmen betrifft es die Identifizierung von Standorten für Antennen zur breitflächigen Netz- und Bereichsabdeckung [30, 31]. Auch im Fachgebiet des Data Minings erfolgt eine Berechnung zentraler Datenobjekte anhand verschiedener Eigenschaften zur Wissensentdeckung [32, 33]. Sogar bei modernen Technologien wie *Software Defined Networking* (SDN) betrifft es die Planung von mehreren SDN-Controllern zur Steuerung von Datenflüssen in Infrastrukturen [34].

Alle diese komplexen Systeme stellen verschiedene Instanzen des k-center und k-cluster Problems dar, wobei ein erhebliches Interesse an effizienten Lösungen durch die breitflächige Anwendbarkeit besteht. Dementsprechend sind stabile und leistungsstarke Optimierungsansätze notwendig. Zudem werden insbesondere im Fall von Notfallszenarien und Hochgeschwindigkeitssystemen verlässliche Ergebnisse in kurzer Zeit benötigt. Im Speziellen betrifft dies Strategien zur Platzierung von den Zentren, zumal sich daran anschließende Planungsschritte und Prozesse darauf abstützen.

Problemstellungen mit einem vergleichbar inhärenten Funktionsprinzip bieten sich an, diese mit einem gemeinsamen Ansatz zu bearbeiten. Daraus ergibt sich die Motivation für einen generischen und einheitlichen Ansatz zur Modellierung dieser Systeme und Szenarien. Die Abbildung 1.9 stellt den ersten Abstraktionsschritt beider Domänen dar. Hierbei übernehmen zentrale Instanzen die Zwischenspeicherung von verschiedenen Verkehrsobjekten auf dem Weg von einer Quelle zu den Nutzern.

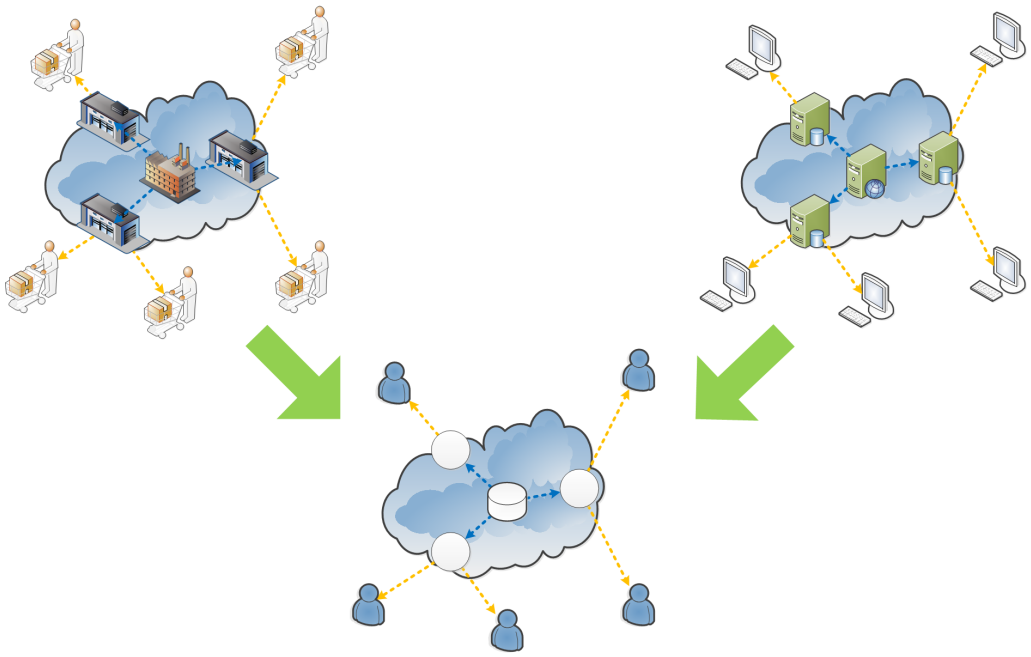


Abbildung 1.9: Generalisierung der spezifischen Transportketten hin zu einer generischen Modellierung unabhängig vom Anwendungsgebiet

Der Vorteil eines generischen Modells besteht in der direkten Nutzung von Lösungsverfahren für mehrere Bereiche. Somit lassen sich Ansätze von einem Anwendungsgebiet in einem anderen übergreifend nutzen. Durch die Vereinheitlichung von Schnittstellen sind Algorithmen und weitere Sachverhalte nur einmalig zu implementieren. Weiterhin vereinfachen auch einheitliche Schnittstellen nach Innen und Außen die Verarbeitung der Daten. Über eine modulare Struktur des Modells wird der Aufwand reduziert domänenabhängige Aspekte in verschiedenen Bereichen zu berücksichtigen. Gleiche Modelleigenschaften sind in einem generischen Ansatz nur einmal umzusetzen, wobei diese dann in mehreren Szenarien nutzbar sind. Zudem ermöglicht dies spezifische Eigenschaften und Nebenbedingungen in gleicher Weise wie bei individuellen Ansätzen einzubeziehen. Darüber hinaus wird die Fähigkeit geschaffen, mehrere Modelleigenschaften gleichzeitig zu betrachten und Zusammenhänge zu analysieren. Somit lassen sich in einem generischen Lösungsansatz gegenseitig beeinflussende Faktoren gezielter betrachten und Sachverhalte übergreifend lösen. Insbesondere betrifft dies kooperative Lösungen zwischen dem Anbieter der Dienste, des Transports und dem Betreiber der Infrastruktur [35, 36]. Letztendlich bietet ein solcher Ansatz die Erschließung neuer Anwendungsgebiete und eine umfassende Betrachtung hinsichtlich einer ganzheitlich optimierten Lösung. Die Herausforderung und Sinnhaftigkeit eines generischen Modells wird auch in den Arbeiten [37–41] dargelegt. Ein Beispiel für eine abstrakte Transportkette ist im Anhang A beschrieben. Dabei sind verschiedene Lösungsmöglichkeiten zur Planung von Zentren analysiert. Es verdeutlicht den Mehrwert hinsichtlich einer strategischen Vorgehensweise. Dies schließt die Entwicklung adaptiver Algorithmen mit ein, welche mittels generischem Modell einem weit gefasstes Spektrum an Problemstellungen zur Verfügung stehen.

1.2.6 Definition von Begriffen

Die Abstraktion von den einzelnen Teilaspekten in den Beispielen erfordert eine übergreifende Beschreibung von Komponenten für einen generischen Ansatz. Für eine einheitliche Semantik wird daher in diesem Abschnitt vorab die Bedeutung von mehreren Begriffen definiert. Die Tabelle 1.1 stellt aus den zwei betrachteten Anwendungsgebieten Logistik und Rechnernetze die wesentlichen Begriffe gegenüber und legt eine generische Bezeichnung fest. Zur Beschreibung des Modellierungskonzeptes und zusätzlicher Sachverhalte werden diese einheitlichen Begriffe in der weiteren Arbeit verwendet.

Bei der generischen Transportkette stellt eine Quelle einen oder mehrere Dienste zur Verfügung. Ein Verteilzentrum, kurz Zentrum, übernimmt jene Dienste in den eigenen Speicher und gibt diese bei Anfrage gegebenenfalls an die Nutzer ab. Die Entitäten des Systems sind dabei über eine Infrastruktur miteinander verbunden.

Neben der Herausforderung einer Modellierung des komplexen Systems erfolgt darüber hinaus eine Lösungsberechnung und Optimierung. Die Bewertung der Systemleistung erfolgt dabei anhand der erreichten Dienstgüte, der sogenannten QoS. Allgemein ist dies durch die Internationale Organisation für Normung (ISO) in der ISO 9000 [42] für Qualitätsmanagementsysteme definiert. Hierbei gibt die Qualität an, in welchem Grad ein Produkt mit den inhärenten Merkmalen bestehende Anforderungen erfüllt. In

Tabelle 1.1: Gegenüberstellung der verschiedenen Begriffe bei Transportketten der Domänen Logistik und Rechnernetze mit entsprechender generischer Bezeichnung

Logistik	Generisch	Rechnernetze
Lagerort, Depot, Verteilstelle	Zentrum	Mirror/Surrogate-Server
Produzent, Hersteller	Quelle	Server, originale Datenbasis
Verbraucher, Konsument	Nutzer	Client, Senke
Produkte, Güter, Ware	Dienst	Daten, Service
Lager	Speicher	Cache
Straßennetz	Infrastruktur	Rechnernetz
Verkehrsweg, Güterfluss	Verbindung	Übertragungsweg, Link

der Logistik entspricht dies dem qualitativen Maß der rechtzeitigen und zuverlässigen Verfügungsstellung von Diensten und Objekten [17]. Somit erfordert ein Transport die Zustellung von Waren in der richtigen Menge, an den vorgesehenen Ort zum spezifizierten Zeitpunkt [43]. Zudem wird der Servicegrad über die Zielfunktionen einer kurzen Reaktionszeit [44] und der geringen Transportzeit definiert [45]. Bei Rechnernetzen ist darüber hinaus QoS durch die *International Telecommunication Union* (ITU) im Standard E.800 [46] und E.802 [47] festgelegt. Diese betreffen die Übertragung digitaler Medien mittels Kommunikationsnetzen und beziehen sich auf Netzvereinbarungen. Hierbei wird QoS als die Gesamtheit aller Eigenschaften eines Systems bezeichnet, das aus dessen Eignung festgelegte Erfordernisse erfüllt. Weiterhin umfasst QoS die Interessen der an dem System beteiligten Entitäten. Die Güte eines Kommunikationsdienstes wird dabei anhand der Parameter Übertragungsbandbreite, Latenzzeiten, Paketverluste und Jitter definiert. Damit eine objektive Bewertung erfolgt, sind die Eigenschaften eines Systems entsprechend zu definieren, sodass diese messbar und vergleichbar sind.

Beide Beschreibungen der Definitionen werden herangezogen und entsprechend der Problemstellung auf allgemeine Systeme von physischen und virtuellen Transportketten adaptiert. In diesem Kontext sind die Interessen der Nutzer, des Systembetreibers und des Besitzers der Infrastruktur zu berücksichtigen. In Anbetracht der beschriebenen Anwendungsgebiete wird QoS für das System der abstrahierten Transportkette wie folgt definiert:

Definition - QoS

Die QoS ergibt sich aus dem Transferweg und der damit einhergehenden Transferzeit, der Reaktionsfähigkeit auf Anfragen von Nutzern an ein Zentrum sowie des Ressourceneinsatzes und der Ressourcenbelastung für das Gesamtsystem in Anbetracht der Interessen der Nutzer und der beteiligten Systembetreiber.

Hierbei besteht das Ziel in der Optimierung der QoS für die Nutzer des Systems, den Systembetreiber und den Infrastrukturbetreiber. Ausschlaggebend dafür ist eine kurze Entfernung zwischen dem jeweiligen Zentren und den zugehörigen Nutzern entsprechend der Verbindungsstrecke entlang der Infrastruktur. Ebenso besteht das Interesse an einer

geringen Fehlerrate bei der Beantwortung von Nutzeranfragen durch das betreffende Zentrum, indem die betreffenden Dienste sofort im Speicher zur Verfügung stehen. Dies hat unter Beachtung der dafür definierten Menge an Zentren und Speicherkapazität zu erfolgen.

1.3 Wissenschaftliche Fragestellungen

Basierend auf der identifizierten Problematik ergeben sich mehrere wissenschaftliche Fragestellungen. Diese werden zunächst genannt und anschließend im Detail erläutert sowie auf Zusammenhänge hingewiesen. Die drei wesentlichen Fragen sind:

1. Was ist ein multikriterieller Optimierungsansatz, der für verschiedene Domänen einer Transportkette anwendbar ist?

2. Wie lassen sich Verhaltensprofile von Nutzern mit in die Planung und Optimierung einbeziehen, um die Leistung des Gesamtsystems zu verbessern?

3. Wie lassen sich verschiedene Optimierungsaspekte der Zentren, welche das Gesamtsystem beeinflussen, für eine integrierte Lösung parallel betrachten?

Auf Grundlage der drei zentralen Fragen dieser Arbeit ergeben sich weiterhin die folgenden Teilfragestellungen. Diese sind in die jeweiligen Themenbereiche eingeteilt.

1. Was ist ein multikriterieller Optimierungsansatz, der für verschiedene Domänen einer Transportkette anwendbar ist?

Für die verschiedenen Anwendungsgebiete ist zu untersuchen, inwieweit diese sich mit einem generischen Ansatz modellieren lassen. Dies erfordert die Betrachtung realer Szenarien unter dem Gesichtspunkt der derzeitigen, spezifischen Lösungsansätze. Dazu sind zunächst Gemeinsamkeiten und Unterschiede in den betrachteten Domänen herauszustellen. Weiterhin ist zwischen domänen-abhängigen und domänen-unabhängigen Anteilen zu unterscheiden. Zur Generalisierung für einen einheitlichen Ansatz sind entsprechende Oberkategorien zu identifizieren und zu abstrahieren, welche sich für die verschiedenen Unterbereiche der Domänen spezifizieren lassen.

Darüber hinaus muss insbesondere untersucht werden, welche bisherigen Ansätze und Lösungsvorschläge für die beschriebene Problemstellung in diesem Forschungsbereich existieren. Hierzu muss zunächst geklärt werden, weshalb diese bestehenden Systeme nicht die abgeleiteten Anforderungen der spezifischen Problemstellung erfüllen. Insbesondere ist hierbei auf den generischen Modellierungsansatz von zweistufigen Transportketten verschiedener Domänen sowie deren Optimierung und Simulation hinsichtlich der szenarienspezifischen Aspekte einzugehen.

Auf dem generischen Modell aufbauend erfolgt anschließend die Berechnung zur Planung von Zentren. Dafür gibt es verschiedene Optimierungsverfahren, die es zu untersuchen gilt. Durch den einheitlichen Modellierungsansatz lässt sich eine Schnittstelle definieren, wodurch die verschiedenen Verfahren sich auf die Problemstellung anwenden lassen. Hierbei ist zu untersuchen, inwieweit sich bestehende Optimierungsalgorithmen als Grundlage für das Anwendungsgebiet verwenden und adaptieren lassen. Eine Analyse der bestehenden Algorithmen ermöglicht die Entwicklung von problemspezifischen Verfahren mit angepassten Lösungen. Ziel ist dabei die Optimierung der Standorte von Zentren in Abhängigkeit der Platzierungsmöglichkeiten und der Infrastruktur. Dies beinhaltet die Analyse, welche Anzahl an Verteilzentren bei optimierter Platzierung für ein Szenario notwendig sind, um einen entsprechenden Grad an Service-Level bestmöglich zu erreichen.

Zur Analyse von Laufzeit-Effekten von solchen komplexen Systemen ist das entstandene Modell mit den geplanten Zentren entsprechend zu simulieren. Dabei ist zu klären, wie dies auf Grundlage des generischen Modells unter Beachtung der domänen-spezifischen Nebenbedingungen zu erfolgen hat.

2. Wie lassen sich Verhaltensprofile von Nutzern mit in die Planung und Optimierung einbeziehen, um die Leistung des Gesamtsystems zu verbessern?

Zur Verbesserung der Gesamtsystemleistung sind die Kenntnisse über die Transportcharakteristiken und das spezifische Nutzerverhalten auszunutzen. Hierbei ist zu analysieren, ob und inwieweit sich die Statistik von vorangegangenen Anfragen bzw. das Interessen-Profil einzelner Nutzer zur Leistungssteigerung verwenden lässt. Darüber hinaus ist zu untersuchen, wie dieses Wissen in die Planung und Optimierung von Zentren mit einbezogen werden kann. Dies beinhaltet die Zuweisung der Nutzer zu den Zentren und das Stellen von Anfragen der Nutzer an das System sowie deren Beantwortung durch ein Zentrum. Eine Analyse dieser Einflussgröße ermöglicht Rückschlüsse auf das Optimierungspotential und den Einfluss historischer Anfragen auf zukünftiges Verhalten. Dabei ist zu klären, welche Informationen für die Entwicklung aufzuzeichnen sind, damit eine effiziente Planung und Optimierung hinsichtlich eines vorteilhaften Systemverhaltens möglich ist.

Für die Beantwortung dieser spezifischen Fragen ist ein Interessens- und Verhaltensprofil von den Nutzer zu modellieren. Dieser Aspekt muss entsprechend in das generische Modell aus der Frage 1 integriert werden. In der Analyse sind die Auswirkungen für das Gesamtsystem unter Beachtung realistischer Gegebenheiten zu ermitteln. Insbesondere ermöglicht dies eine Bewertung der notwendigen Speicherkapazität der Zentren für eine effiziente Funktionsweise des Gesamtsystems. Darüber hinaus ist das Potential der Verwaltungsstrategien bei den Zwischenspeichern auszunutzen, sodass Ressourcen adäquat verwendet werden.

3. Wie lassen sich verschiedene Optimierungsaspekte der Zentren, welche das Gesamtsystem beeinflussen, für eine integrierte Lösung parallel betrachten?

Ziel dieser Frage ist die gleichzeitige, parallele Betrachtung der verschiedenen Optimierungsaspekte, welche das Gesamtsystem beeinflussen. Für eine ganzheitliche Lösung sind die Zentren unter Berücksichtigung der Domäneneigenschaften und mehrerer Einflussgrößen zu optimieren und zu dimensionieren. Dies erfordert multikriterielle Optimierungsansätze beziehungsweise eine Abbildung der Optimierungsziele auf ein kombiniertes Kriterium. Eine Formulierung multidimensionaler Zielstellungen entsteht durch das Vorliegen eines Zielkonfliktes. Hierbei führt in aller Regel eine Verbesserung hinsichtlich einer Zielsetzung zu einer Verschlechterung in mindestens einem anderen Zielkriterium [48]. Demzufolge existieren bei multikriteriellen Problemstellungen keine alternativen Zielfunktionen, welche für alle Zielkriterien gleichermaßen optimal sind.

Des Weiteren ist zu untersuchen, ob ein generischer Ansatz umsetzbar ist und inwieweit sich damit die Problemstellungen ähnlichen Funktionsprinzips spezifisch lösen lassen. Dabei ist die Anwendbarkeit von Lösungsansätzen eines Anwendungsgebietes auf ein anderes zu untersuchen. Die Herausforderungen bei der Optimierung bestehen in der Identifizierung von vorteilhaften Standorten unter Berücksichtigung der Nebenbedingungen der verschiedenen Domänen und mehrerer Zielkriterien. Bei der Planung der Zentren ist neben der Entscheidung über die Standorte auch die Zuordnung der Nutzer zu bedenken. Die Optimierung der Zuordnung hat unter Berücksichtigung der Interessens- und Verhaltensprofile zu erfolgen, sodass die spezifischen Domäneneigenschaften vorteilhaft in Bezug des einzelnen Nutzers und des Gesamtsystemverhaltens genutzt werden. Insbesondere stellt sich hierbei die Frage, wie sich das Wissen der Nutzerprofile in die Standortwahl und die Zuordnung von Nutzern zu Zentren mit in die Optimierung einbeziehen lässt. In Abhängigkeit der Domänen sind für das Systemverhalten Strategien zur Verteilung von Diensten sowie deren Ersetzung im Zwischenspeicher zu analysieren.

1.4 Zielstellung der Arbeit

Das Ziel der Arbeit besteht in der Entwicklung eines generischen Modells für verschiedene Anwendungsgebiete von zweistufigen Transportketten. Für die dadurch abgebildeten Problemstellungen werden Lösungen mittels Optimierungsalgorithmen gegeben. Dies hat fallspezifisch zu erfolgen. Zudem ist das Modell hinsichtlich einer multikriteriellen Optimierung und Analyse ganzheitlicher Aspekte des Systems auszulegen. Die Optimierung erfolgt dabei mit der Zielsetzung die QoS für den Nutzer und den Systembetreiber zu verbessern. Dazu ist das Szenario hinsichtlich der operativen Ressourcenauslastung von Infrastruktur, Zentren und des zu erwartenden Nutzerverhaltens zu optimieren. Das entstehende System wird mittels Simulation evaluiert. Hierbei hat die Simulation der Transportkette unter Beachtung von Nebenbedingungen der Domäne zu erfolgen. Zur Leistungsverbesserung werden die Interessen-Profile der Nutzer in die Optimierung mit einbezogen, sodass die Kenntnis darüber sich vorteilhaft auf das Systemverhalten auswirkt. Hierbei erfolgt die Berechnung der sogenannten Pareto-Front. Diese besteht aus einer Menge an Lösungen, welche einen optimalen Zustand besitzen.

Bei solch einer Lösung ist es nicht möglich, ein Zielkriterium zu verbessern, ohne zugleich ein anderes Kriterium zu verschlechtern. Dieser Zustand ist hinsichtlich mehrerer Kriterien Pareto-Effizient. Da die Platzierung von den Zentren die Lösungsqualität maßgeblich beeinflusst, wird dafür ein angepasstes Optimierungsverfahren entwickelt. Ein weiteres Ziel der Arbeit besteht in der bedarfsgerechten Dimensionierung der Zentren hinsichtlich Anzahl, Standort, Kapazität und Speicherstrategie. Der generische Ansatz ist anhand von Szenarien aus den Bereichen des Logistik und Rechnernetze zu evaluieren. Zudem wird die weitere Anwendungsmöglichkeit im Rahmen der Geolokalisation gezeigt, um Messpunkte zur Standortbestimmung zu ermitteln.

1.5 Aufbau der Dissertation

Die Abbildung 1.10 zeigt die grundlegende Struktur der vorliegenden Arbeit und die inhaltliche Vorgehensweise bei der Beschreibung des Sachverhaltes. Zu Beginn erfolgte eine genauere Betrachtung des Wissenschaftsgebiets der Transportnetze, welche anschließend in Bezug zur Logistik mit dem FLP und Rechnernetze mit CDN dargestellt wird. Aus den Bereichen ergibt sich die Motivation für einen generischen Ansatz. Dabei ergeben sich vielfältige wissenschaftliche Fragestellungen im Kapitel 1.3. Diese bilden die Grundlage zur Formulierung der Zielsetzung der Dissertation im Kapitel 1.4.

Die ausführliche Betrachtung der Problemstellung erfolgt im Kapitel 2. Diese unterteilt sich gemäß der wissenschaftlichen Fragestellungen in drei Teilbereiche. Zunächst erfolgt eine Erläuterung, welche Herausforderungen bei der Entwicklung des generischen Modells zu bewältigen sind. Darauf aufbauend wird die Simulation der Prozessabläufe unter Einbeziehung der Verhaltensweisen näher analysiert. Anschließend sind die Probleme hinsichtlich einer multikriteriell optimierten Lösung im Detail aufgeführt.

Zur Problemlösung werden in Kapitel 3 mehrere Anforderungen aufgestellt und anschließend im Kapitel 4 der Stand der Wissenschaft umfassend recherchiert und analysiert.

Im Hauptteil der Arbeit wird das generische Modell entwickelt. Dieses Kapitel 5 zeigt darüber hinaus die Anwendbarkeit für die Bereiche Logistik mit dem FLP, Rechnernetze mit CDN und Messpunktbestimmung zur Geolokalisation. Zudem wird ein neuer Algorithmus für die Optimierung der Standortplanung vorgeschlagen.

In Kapitel 6 wird die Tragfähigkeit des Konzeptes mittels der Umsetzung als Analysewerkzeug zur Lösungsberechnung sowie umfangreichen Analysen zur Optimierung und Simulation von verschiedenen Problemstellungen im Bereich der Transportketten beschrieben. Dies beinhaltet Tests zur Leistungsfähigkeit des entwickelten Optimierungsalgorithmus bei der Platzierung. Weiterhin ist der spezifische Ansatz der Geolokalisation weiteren Methoden gegenübergestellt.

Abschließend erfolgt in der Evaluation und Auswertung eine Diskussion der Ergebnisse mit einer Zusammenfassung und Ausblick der vorliegenden Arbeit.

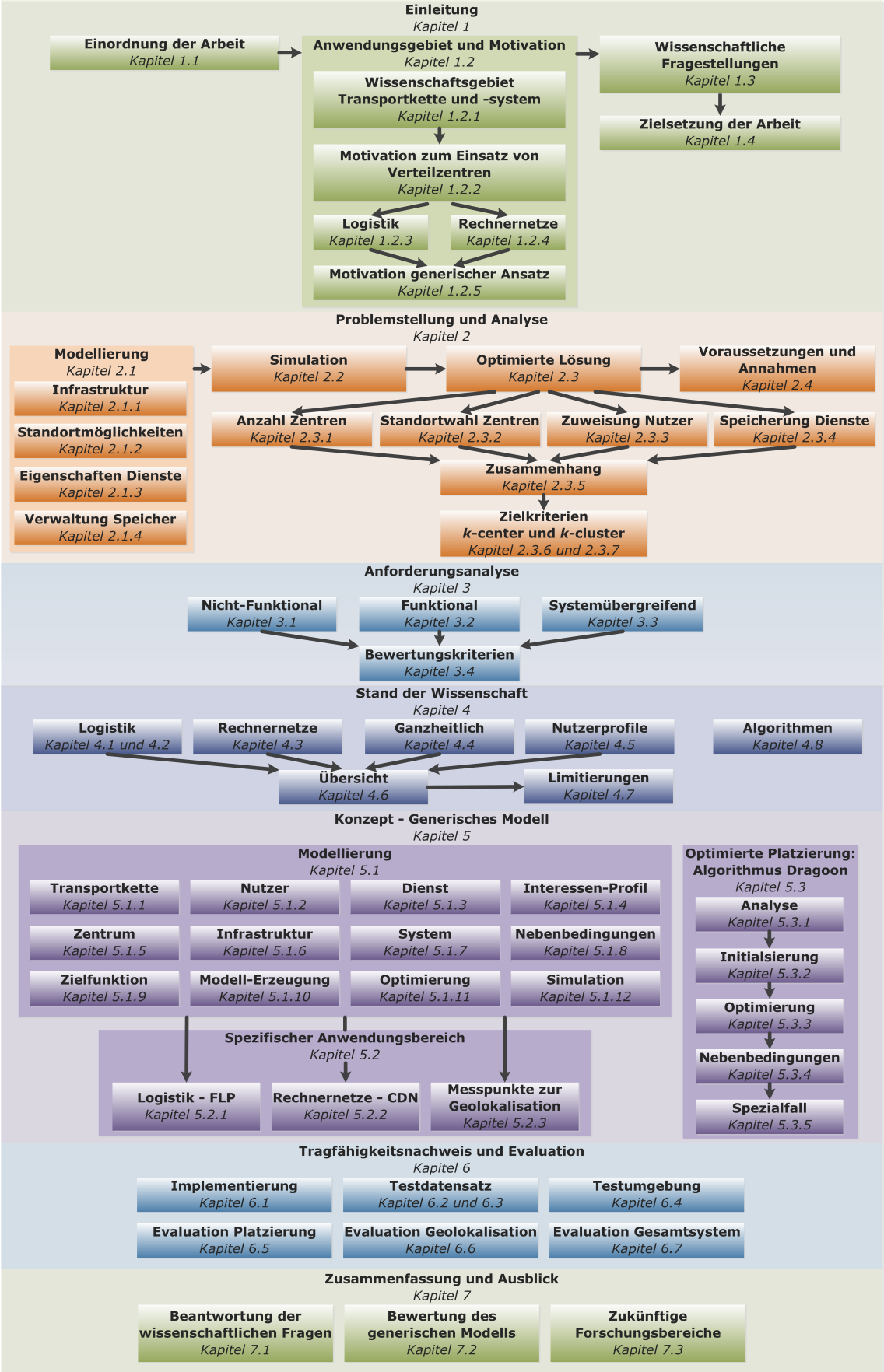


Abbildung 1.10: Aufbau und Vorgehensweise bei der Beschreibung der vorliegenden Arbeit

2 Problemstellung

Die in dieser Arbeit betrachtete Problemstellung besteht in der generischen Modellierung, Simulation und Optimierung von zweistufigen Transportketten. Unter Verwendung der Eingangsdaten eines gegebenen Szenarios ist ein angepasstes Modell auf Basis des generischen Ansatzes zu erstellen. Auf Grundlage dieses Modells hat anschließend die Berechnung einer optimierten Lösung zu erfolgen. Die Ergebnisse und das Modell sind in Simulationen zu analysieren und zu überprüfen. Die detaillierte Beschreibung der Problemstellung ist in diese drei Themenbereiche aufgeteilt, auf die im Folgenden näher eingegangen wird.

2.1 Generische Modellierung

Bei der Modellierung sind Szenarien von zweistufigen Transportketten abzubilden. Zur Minimierung des zukünftigen Entwicklungsaufwandes und gemäß der Vorteile im Abschnitt 1.2.5 ist dies mit einem gemeinsamen und einheitlichen Ansatz durchzuführen. Das entstehende Modell muss dazu flexibel für verschiedene Anwendungsbereiche anwendbar sein, insbesondere mit dem Schwerpunkt der betrachteten Domänen Logistik und Rechnernetze.

Die Abbildung 2.1 verdeutlicht die Problemstellung der Entwicklung eines generischen Modells. Der dabei entstehende einheitliche Ansatz bildet die Basis für weitere Berechnungen wie unter anderem der optimierten Lösungsfindung und der simulativen Analyse.

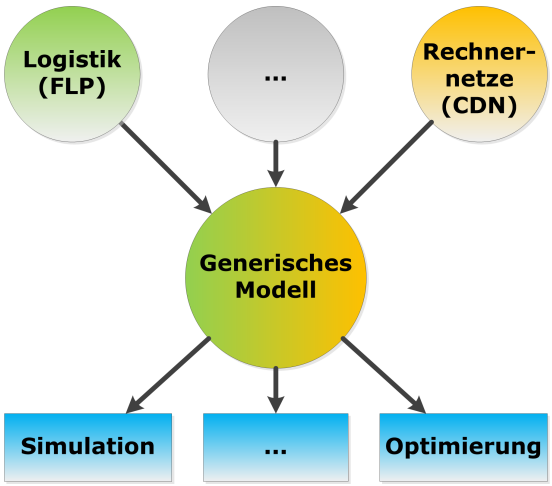


Abbildung 2.1: Problem der Entwicklung eines generischen Modells für mehrere Anwendungsgebiete

Bei der Modellierung sind mehrere Aspekte zu berücksichtigen. Die Zentren der geplanten Transportkette müssen geographisch platziert und in deren Eigenschaften geplant werden. Zur Versorgung der vorgegebenen Nutzer erfolgt im Modell eine passende Zuweisung zu diesen Zentren. Die Nutzer weisen hierbei ein spezifisches Interessen-Profil auf, das in der Planung und Lösungsfindung zu berücksichtigen ist. Dies stellt insbesondere im Zusammenhang mit dem Nutzerverhalten für die Modellierung eine Herausforderung dar. Weiterhin sind die verschiedenen Prozessabläufe sowohl von der Transportkette als auch eines Zentrums nachzubilden. Dazu gehört das Anfordern von Diensten der Nutzer am Zentrum und die Bereitstellung durch das System aus Zentren des Transportnetzes. Ein Zentrum muss dabei die vorhandenen Dienste im Speicher beachten und bei Bedarf weitere Prozesse anstoßen. Für eine zufriedenstellende Beantwortung hat das jeweilige Zentrum den entsprechenden Dienst selbst im Speicher vorzuweisen. Somit muss das Modell neben der Abbildung der Teilkomponenten des Systems und deren Funktionalität auch übergreifende Prozesse ermöglichen. Ein weiteres Problem besteht dabei in der Sicherstellung der Erweiterbarkeit der Teilkomponenten, um zusätzliche Eigenschaften spezifischer Sachverhalte bei einer Weiterentwicklung mit abzubilden. Zudem ist das Modell trotz des komplexen Sachverhaltes an die unterschiedlichen Ausprägungen einer Transportkette in den Anwendungsgebieten flexibel anpassbar zu gestalten. Darüber hinaus sind bei der generischen Modellierung die domänenspezifischen Aspekte mit zu berücksichtigen, um weitere Einflussfaktoren mit einzubeziehen.

Die wesentlichen domänenspezifischen Aspekte für die Lösungsfindung und Optimierung sind:

- Infrastruktur und Vernetzung der Teilnehmer einer Transportkette
- Standortmöglichkeiten für Zentren
- Eigenschaften der Dienste
- Speicherung und Verwaltung im Zentrum

2.1.1 Infrastruktur und Vernetzung der Teilnehmer einer Transportkette

Die Beachtung der Eigenschaft Infrastruktur ist von maßgeblicher Bedeutung für das Modell, da diese die Teilnehmer der Transportkette miteinander verbindet und die Grundlage für die weitere Planung darstellt. Für jedes Anwendungsgebiet und in Abhängigkeit des Szenarios existiert eine spezifische Infrastruktur, welche individuell bei der Planung mit einzubeziehen ist. Die Berechnung der Entfernung zwischen zwei Teilnehmern hat auf Grundlage dieser zu erfolgen. Das generische Modell muss sich dabei unabhängig von der definierten Infrastruktur erstellen lassen. Über die Infrastruktur ist jeder Teilnehmer mit jedem verbunden, sodass die Erreichbarkeit sichergestellt wird und die Bereitstellung der Dienste erfolgen kann. Hierbei ist zwischen direkten und indirekten Verbindungen zu unterscheiden. In Abhängigkeit des Szenarios ist die Infrastruktur entweder explizit vorgegeben oder wird implizit über eine Entfernungsberechnung spezifiziert. Bei expliziter Vorgabe muss zur Bestimmung der Entfernung zunächst der entsprechende Verbindungspfad bestimmt werden, bevor eine Berechnung der Entfernung

erfolgt. Im Fall einer impliziten Angabe geschieht dies durch die Entfernungsberechnung, sodass standardmäßig von einer Vollvermaschung mit direkten Verbindungen zwischen den Teilnehmer auszugehen ist.

2.1.2 Standortmöglichkeiten für Zentren

Bei der Standort-Auswahl von Zentren sind bei der Modellierung und Planung die verschiedenen Platzierungsmöglichkeiten zu beachten. In Anbetracht der unterschiedlichen Anwendungsgebiete ergeben sich dabei die Nebenbedingungen, Knotenplatzierung und freie Platzierung [49–51]. Bei der Knotenplatzierung werden spezifische Möglichkeiten für die Standorte von Zentren vorgegeben. Die Platzierung eines Zentrums hat somit an einer vordefinierten Position zu erfolgen. Dies ist unter anderem bei der Vorgabe einer Infrastruktur bedingt. Dem gegenüber steht die freie Platzierung. Hierbei wird bei einem Szenario keine Einschränkung in der Standortwahl vorgegeben. Weiterhin gibt es beliebige Mischformen, bei denen eine freie Platzierung auf bestimmte Regionen beschränkt ist. Das Problem besteht in der Abbildung der verschiedenen Standortmöglichkeiten im Modell über eine einheitliche Schnittstelle zur Optimierung.

2.1.3 Eigenschaften der Dienste

Die von einem Nutzer angeforderten Dienste besitzen unterschiedliche Eigenschaften in Abhängigkeit des Anwendungsgebietes. Im Bereich von FLP handelt es sich um Dienste, die in stückweisen Mengen transportiert werden. Bei der Abgabe von einer Quelle oder einem Zentrum an einen Nutzer steht der Dienst für andere Nutzer nicht mehr zur Verfügung und wird somit verbraucht. In der Domäne von CDN lässt sich im Gegensatz dazu ein Dienst durch Kopieren vervielfältigen. Nach der Beantwortung einer Anfrage steht der jeweilige Dienst somit bei der Quelle immer noch für andere Nutzer zur Verfügung. Beide Ausprägungen der Eigenschaft müssen durch den gemeinsamen Ansatz abgedeckt werden.

2.1.4 Verwaltung und Speicherung der Dienste im Zentrum

Zur Analyse von operativen Laufzeiteffekten sind die internen Abläufe sowie die externen Interaktionen bei einem Zentrum zu modellieren und zu simulieren. Dies schließt die Initiierung der Anfragen über die Verarbeitung bei der Dienstübertragung und das Vorgehen bei der Zwischenspeicherung bis hin zur Beantwortung mit ein. Im Bereich von FLP sind die physischen Dienste vorab im Zentrum zu lagern, damit diese bei Anfrage sofort zur Verfügung stehen. Im Vergleich zur Domäne CDN lassen sich die virtuellen Dienste bei Bedarf zeitnah von einer Quelle holen, sodass eine Anforderung nicht langfristig im Voraus geplant werden muss. Entsprechend der Domänenspezifika hat diese proaktive bzw. reaktive Speicherung von Diensten und die Verwaltung des Speichers im Zentrum zu erfolgen. Das generische Modell ist für beide Vorgehensweisen auszulegen.

2.2 Simulation des generischen Modells

Neben der Modellierung der Transportkette hat die Simulation der Prozessabläufe zu erfolgen, um über die statische Lösungsfindung und Optimierung hinaus Aussagen zu treffen. Zur Bewertung des Systemverhaltens sind dazu weitere dynamische Einflüsse und Reaktionsweisen mit in die Betrachtung einzubeziehen. Für ein Entscheidungsunterstützungssystem bedarf es der Analyse von stochastischen Einflüssen auf das entwickelte Modell des Transportsystems eines Szenarios. Daher wird die Aufgabenstellung unter Beachtung zufälliger Ereignisse und wahrscheinlichkeitsbedingten Entscheidungen während der Laufzeit des Systems betrachtet. Das Problem besteht im Testen der Auswirkungen von verschiedenen Verhaltensweisen und den Entschlüssen von der Planung auf Grundlage des generischen Ansatzes. Insbesondere ist es dazu notwendig, die bei der Lösungsfindung und Optimierung erfolgten Zwischenergebnisse zu überprüfen und auszuwerten. Mit einer Simulation der Prozesse des Transportsystems lassen sich quantitative Aussagen über eine Lösung für ein Szenario treffen. Eine umfangreiche Überprüfung von alternativen Vorschlägen und unterschiedlichem Systemverhalten in Bezug auf die Zentren ist auf Grund der Anzahl von Möglichkeiten problematisch.

Da ein Zentrum nur eine begrenzte Menge an Diensten zu einem Zeitpunkt bereitstellen kann, erfolgt die Speicherung dynamisch. Der Speicher ist dazu mittels einer Verwaltungsstrategie zu organisieren. Somit hat auch die Größe des Speichers und das Verhalten des Zentrums einen Einfluss auf die Zuweisung der Nutzer. In Abhängigkeit des Anwendungsgebietes bestehen hierfür mehrere Möglichkeiten. Das Problem besteht in der Analyse von operativen Laufzeiteffekten und der angepassten Ersetzungsstrategie, da dies von den Anfragen der Nutzer und deren Interessen abhängt. Hierbei müssen die internen Abläufe in einem Zentrum modelliert und simuliert werden. Insbesondere gehört dazu die proaktive bzw. reaktive Speicherung von Diensten.

2.3 Optimierte Lösungsfindung

Für das Management des betreibenden Unternehmens ergeben sich mehrere Herausforderungen bei der Planung und Optimierung eines komplexen Transportsystems. Das Problem besteht in der Anwendbarkeit der Lösungsalgorithmen, sodass diese bei den unterschiedlichen Ausprägungen der Nebenbedingungen einsetzbar sind. Auf der Grundlage des generischen Modellierungsansatzes hat die Berechnung einer Lösung und die Optimierung zu erfolgen. Die vier wesentlichen Entscheidungen betreffen die Anzahl der Zentren im System, die Wahl der Standorte von den Zentren sowie die Zuweisung aller Nutzer zu deren primären Zentrum und der Speicherung von Diensten im jeweiligen Zentrum. Dazu werden die einzelnen Merkmale zunächst genauer erläutert, bevor das Problem der Optimierung beschrieben wird.

2.3.1 Anzahl der Zentren im System

Bevor eine Planung und Platzierung der Zentren erfolgen kann, muss eine Entscheidung über die Anzahl getroffen werden. Mit steigender Anzahl erhöht sich die Menge an notwendigen, einzusetzenden Ressourcen, ermöglicht jedoch auch mehr Flexibilität bei der Planung der Zentren und der Versorgung der Nutzer. Dabei stellt sich die Frage, mit welcher Anzahl von Zentren wird für ein Szenario welches Level von QoS erreicht. Da dies vorab auf Grund der Komplexität des Szenarios nicht überschaubar ist, stellt die Ermittlung einer vorteilhaften Menge an Zentren ein Problem dar.

2.3.2 Standortwahl der Zentren

Nach der Festlegung der Anzahl von Zentren ergibt sich die Fragestellung der Standortwahl, insbesondere wo ein Zentrum bei Beachtung der Infrastruktur zu platzieren ist. Die Entscheidung wird dabei von mehreren Bedingungen beeinflusst. Da die Zentren für die Nutzer in der Transportkette zuständig sind, werden die Nutzer jeweils einem primären Zentrum zugewiesen. Dieses ist für eine effiziente Versorgung der zugeteilten Nutzer zuständig. Die Zuweisung erfolgt auf Grund von mehreren Kriterien. Vorrangig sind die Zentren mit kurzer Entfernung zu den Nutzern zu planen, um eine Verbesserung der QoS und Reduktion der Infrastruktur-Last zu erreichen [52]. Neben der Entfernung eines Nutzers zu einem Zentrum haben unter anderem die Interessen-Profile der Nutzer Auswirkungen auf die Zuweisung. Dadurch beeinflussen die einem Zentrum zugewiesenen Nutzer während der Lösungsoptimierung dessen Standort. Somit entsteht ein multikriterielles Problem, welches die Wahl der Standorte beeinflusst und sich damit auf die optimierte Platzierung auswirkt.

2.3.3 Zuweisung der Nutzer zu einem Zentrum

Auf Grund der Zwischenspeicherung von Diensten in einem Zentrum ist zu überdenken, welcher Nutzer welches Zentrum vorrangig anzufragen hat. Das nächstgelegene Zentrum stellt hierbei nicht unbedingt die beste Option dar, um das Nutzungsprofil zu erfüllen. Für die effektive Bereitstellung eines vom Nutzer angeforderten Dienstes muss das Zentrum diesen auch zur Verfügung haben. Hierbei ist zu analysieren, welche Dienste an den jeweiligen Standorten bereitzustellen sind. Das Problem der passenden Zuweisung eines Nutzer zu einem geplanten Zentrum ist vor allem von der Wahl des Standortes beeinflusst und somit von der Platzierung des Zentrums abhängig.

2.3.4 Speicherung der Dienste in den Zentren

Zur Beantwortung der Nutzeranfragen durch ein Zentrum muss dieses die Dienste in seinem eigenen Speicher vorhalten. Auf Grund der begrenzenden Ressourcen ist ein Zentrum nicht in der Lage alle in der Domäne existierenden Dienste dauerhaft vorzuhalten. Daher besteht das Problem in der Entscheidung über die durch das Zentrum angebotenen Dienste. Weiterhin ist bei der Verwaltung des Speichers über die Ein- und Auslagerung

von Diensten unter Beachtung der maximalen Kapazität zu urteilen. Bei der Planung der Zentren ist die Größe des Speichers zu beschließen und entsprechend so zu planen, dass eine effiziente Funktionsweise ermöglicht wird. Dadurch lässt sich das operative Verhalten und die sofortige Bereitstellung der Dienste verbessern.

2.3.5 Zusammenhang der Optimierungsaspekte

Die vier Teilprobleme haben einen direkten und indirekten wechselseitigen Einfluss aufeinander, wodurch sich ein komplexes Optimierungsproblem ergibt. Der Zusammenhang basiert auf der Optimierung des Transportweges sowie der Anfrage-Zuweisung unter Beachtung der Vorgaben. Das Ziel der Optimierung besteht in der Erhöhung der QoS für den Nutzer als auch der zur Beantwortung von Anfragen notwendigen Ressourcen. Der Zusammenhang der vier wesentlichen Teilprobleme ist in der Abbildung 2.2 dargestellt.

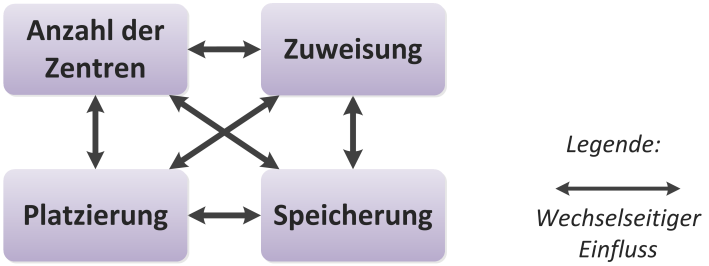


Abbildung 2.2: Wechselseitiger Einfluss der vier Optimierungsaspekte

Die Platzierung von Zentren und die Identifizierung von zentralen Standorten basiert auf dem *k*-center Problem. Die Zuweisung von Nutzern zu einem passenden Zentrum ist vergleichbar mit dem *k*-cluster Problem. Beide Teilprobleme sind für sich nachweislich NP-schwer [53–55], siehe Anhang B. Diese müssen zur Berechnung einer ganzheitlichen Lösung für die umfassende Problemstellung in Kombination betrachtet werden, da diese sich gegenseitig beeinflussen [56, 57]. Mit geänderter Anzahl an Zentren ergibt sich auf Grund der Komplexität des Problems ein anderer Sachverhalt für die Optimierung, sodass dies mit einer anderen Aufgabenstellung gleichzusetzen ist und zu einer vollständigen erneuten Planung und Optimierung führt. Mit angepasster Speicherkapazität der Zentren ergibt sich bei der Bereitstellung von Diensten ein geänderter Sachverhalt bei der Beantwortung der Anfragen, wodurch die anderen Teilprobleme und diesbezüglichen Beurteilungen beeinflusst werden. Die Entscheidung hinsichtlich eines optimalen Speicherverhaltens ist ebenfalls einem NP-Problem gleichzusetzen [58, 59].

Der folgende Konflikt-Kreis beschreibt den Zusammenhang der Optimierungskriterien allgemein und verdeutlicht das Optimierungsproblem. Es zeigt das Potential zur Systemverbesserung.

Konflikt-Kreis:
Eine höhere Anzahl von Zentren führt zu einer verbesserten Platzierung mit kürzeren Entfernungen und feingranulareren Gruppierung. Dadurch verringert sich die durchschnittliche Strecke eines Nutzers zu den Zentren und die QoS wird erhöht. Nutzer erhalten dadurch mehr näher gelegene Zentren und dabei die Möglichkeit von verschiedenen

Zuweisungen. Ein einzelnes Zentrum bekommt weniger Anfragen durch die Verteilung auf die Zentren. Die Notwendigkeit für eine große Kapazität der Zwischenspeicher wird reduziert und die Anfragen nach gleichen Diensten an einem Zentrum verringern sich. Dies erhöht die Fehlerrate bei der Zwischenspeicherung und verringert die QoS. Dies lässt sich im Gegenzug mit einer reduzierten Anzahl von Zentren und angepasster Zuweisung verbessern sowie mit mehr zusammengefassten Nutzern.

2.3.6 Platzierung gemäß dem *k*-center und *k*-median Problem

Die Aufgabenstellung der Platzierung von Zentren und der Identifizierung von mittig gelegenen Standorten bezieht sich in den zuvor beschriebenen Szenarien auf die Entfernung zwischen einem Nutzer und seinem zugewiesenen Zentrum. In Abhängigkeit der Zielfunktion für das gesamte Szenario basiert die Platzierung auf dem mathematischen *k*-center bzw. *k*-median Problem. Beim *k*-center Problem besteht das Ziel den maximalen Zielfunktionswert von den Nutzern zu minimieren [60, 61]. Ausschlaggebend ist somit der Nutzer, welcher die größte Entfernung zu seinem zugewiesenen Zentrum hat. Beim *k*-median Problem besteht dagegen das Ziel den Durchschnitt von den Zielfunktionswerten aller Nutzer zu minimieren [62]. Hierbei wird die durchschnittliche Entfernung von allen Nutzern zu deren jeweiligen Zentrum betrachtet. Daher werden beim *k*-center und *k*-median Problem die Nutzer jeweils zu deren nächst gelegenen Zentrum zugewiesen. Die Abbildung 2.3 zeigt den Unterschied in der Platzierung von den beiden Aufgabenstellungen anhand eines Szenarios mit 150 Knoten. Die blaue Linie gibt die Zuweisung der grünen Knoten zu den roten Zentren an. Durch die Berücksichtigung weiterer Nebenbedingungen in den Anwendungsgebieten aus Abschnitt 1.2 wird von dieser Vorgehensweise zu Gunsten weiterer Leistungsmerkmale bei der Zuweisung abgewichen. Insbesondere durch die Beachtung stochastischer Entscheidungen und zufälliger Einflüsse verkompliziert sich das Problem.

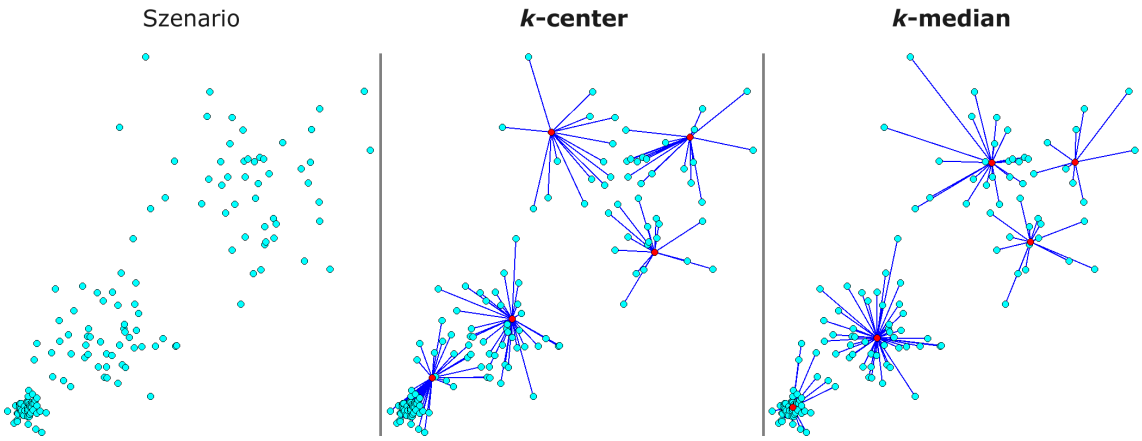


Abbildung 2.3: Gegenüberstellung einer optimierten Lösung zu dem *k*-center und *k*-median Problem mit der Platzierung von fünf Zentren

Das statische Problem der Platzierung ohne Kapazitätsbeschränkung der Zentren wurde erstmals im Anwendungsbereich Logistik mit dem FLP beschrieben [63, 64]. Dabei können einem Zentrum beliebig viele Nutzer zugeordnet sein. Das Problem lässt sich durch einen

zusammenhängenden und ungerichteten Graphen $G(V,E)$ beschreiben. Dieser Graph G besteht aus einer Menge von Knoten ($v_i \in V$ *mit* $i := \{1,..,n\}$) und einer Menge von Kanten ($e_q \in E$ *mit* $q := \{1,..,m\}$). Für eine vorgegebene Menge V von Standorten ist eine vordefinierte Anzahl K (*mit* $0 < k \leq |V|$ *und* $k \in \mathbb{N}$) von zentralen Positionen zu bestimmen. In Bezug zum betrachteten Anwendungsgebiet können die Knoten und Kanten unterschiedliche Wertigkeiten in direkter Abhängigkeit von deren Eigenschaften haben. Die Zielfunktion d definiert den Zielfunktionswert $d(v_i, v_j)$ für eine Kante $e(v_i, v_j)$ zwischen zwei direkt miteinander verbundenen Knoten (v_i, v_j). Für eine Verbindung von zwei Knoten $v_{i,j}$ über mehrere Kanten $e_{s,..,t}$ hinweg werden die Zielfunktionswerte $d_{s,..,t}$ entsprechend der Teilverbindungen aufsummiert. Die Menge aller Zielfunktionswerte $d(v_i)$ von den Knoten ist als D definiert. Der Graph genügt den Anforderungen der Vollständigkeit und erfüllt die Dreiecksungleichung 2.1 mit $v_i, v_j, v_t \in$ V.

$$d(v_i, v_j) \leq d(v_i, v_t) + d(v_t, v_j) \tag{2.1}$$

Zur Evaluierung eines Szenarios verwendet dazu die Zielfunktion d die beste Verbindung $e_{s,..,t}$ von einem Knoten v_i zu einem der berechneten Standorte für ein Zentrum ($k_u \in K$ *mit* $u := \{1,..,l\}$ *und* $l \leq n$).

Die Menge der Zentren K bildet somit eine Hülle für die Knoten V, wenn diese mindestens innerhalb der Entfernung eines Zentrums von höchstens r erreichbar ist. Dies ist der Fall, wenn d(v,K) $\geq$ r für alle Knoten v $\in$ V gilt. Der minimale Wert für r, welcher für die Menge K eine Hülle bildet, wird als Radius r(K) der Hülle von K bezeichnet [65]. Das Optimierungsziel besteht hierbei eine vorgegebene Anzahl von höchstens k Zentren so zu platzieren, dass für die Zielfunktion der bestmögliche Wert erreicht wird [66, 67]. In Abhängigkeit der Platzierungsbedingung ist dabei ein Zentrum k_u an einem Knoten zu platzieren oder lässt sich entsprechend der Nebenbedingungen frei positionieren.

Die Zielfunktion des k-center Problems ist durch die Gleichung 2.2 beschrieben.

$$D_{center}(K) = \min \max_{v_i=1,...,n} \min_{k_u \in K} d(v_i, k_u) \tag{2.2}$$

Die Zielfunktion des k-median Problems ist durch die Gleichung 2.3 beschrieben.

$$D_{median}(K) = \min \sum_{v_i=1}^{n} \min_{k_u \in K} d(v_i, k_u) \tag{2.3}$$

2.3.7 Gruppierung gemäß dem *k*-cluster Problem

Neben der Berücksichtigung der Entfernung stellt eine passende Zuweisung entsprechend des Interessen-Profils eines Nutzer zu einem Zentrum ein Problem dar [68, 69]. Für eine zügige Bereitstellung der angeforderten Dienste muss das angefragte Zentrum den Dienst im Speicher haben. Da sich ein Zentrum nach den Interessen der Nutzergruppe richtet, wird durch die Anfragen eines einzelnen Nutzers mitbestimmt, welche Dienste in einem Zentrum zwischengespeichert werden. Auf Grund von begrenzter Speicherkapazität ist ein Zentrum in der Lage, nur einen Teil aller Dienste temporär vorzuhalten. Durch

eine entsprechende Gruppierung und Zuweisung der Nutzer zu den einzelnen Zentren lassen sich diese vorteilhaft ansteuern und die notwendige Speicherkapazität minimieren. Hintergrund ist die Erhöhung der positiven Korrelation von Anfragen an den zentralen Knotenpunkten. Das Problem besteht somit in der Identifizierung von Nutzer-Gruppen, deren gruppierte Interessen-Profile zu einer effizienten Nutzung des Zentrums führen. Dies entspricht dem k-cluster Problem, bei dem Knotenpunkte anhand ähnlicher Eigenschaften in Gruppen zusammengefasst werden. Die Abbildung 2.4 verdeutlicht die Problemstellung der Gruppierung. Auf der linken Seite stehen die verschiedenen Farben für unterschiedliche Interessen-Profile. Bei der rechten Seite stellen die Farben die identifizierten Gruppierungen dar.

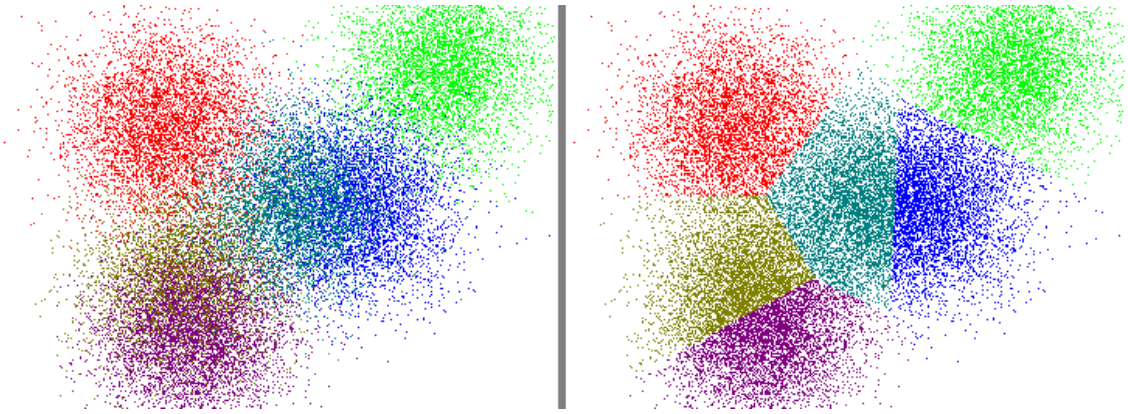


Abbildung 2.4: Beispiel des k-cluster Problems mit originalen Daten links und dem Ergebnis rechts (Vgl. [70])

In den betrachteten Anwendungsgebieten besitzt dabei jeder Nutzer v_i ein Interessen-Profil pf_i, nachdem sich dieser entsprechend verhält und Anfragen an ein Zentrum stellt. Das Interessen-Profil setzt sich dabei aus verschiedenen Eigenschaften zusammen. Es ist davon auszugehen, dass die Interessen-Profile der Nutzer dabei individuell sind und die Eigenschaften demzufolge eine unterschiedlich starke ausgeprägt haben.

Zur Verbesserung der Prozesse in einem Zentrum und der Dienstbereitstellung ist die Kenntnis der Nutzerinteressen für die Optimierung zu verwenden. Eine passende Zuweisung vermeidet das Weiterleiten von Anfragen an andere Zentren und minimiert die Last auf die Infrastruktur. Zu Gunsten einer verbesserten Zuweisung wird eine höhere Entfernung zwischen Nutzer und Zentrum akzeptiert.

Die Gleichung 2.4 beschreibt die Zielfunktion $C_{cluster}$ für die Gruppierung der Nutzer. Dabei erfolgt eine Korrelation der Interessen-Profile $pf_i \in P \;\; mit \;\; i := 1,..,n)$ für das gesamte System ohne Abbildung der operativen Auswirkungen bei den Zentren. Die Korrelationsfunktion ρ definiert den Korrelationskoeffizienten $\rho(pf_i|pf_j)$ für zwei Profile (pf_i, pf_j). In dem betrachteten Szenario geschieht dies paarweise für jeden Nutzer mit dem zugehörigen Zentrum. Für eine Gruppe von mehreren Nutzern werden von den Korrelationskoeffizienten $\rho(pf_i, ..., pf_j)$ entsprechend aufsummiert. Die Menge aller Korrelationskoeffizienten $\rho(pf_i)$ von einem Nutzer ist als C definiert.

$$C_{cluster} = \max \sum_{v_i=1}^{n} \rho(pf_i | pf_{k(v_i)}) \tag{2.4}$$

Die Problemstellung der Platzierung wird durch die Berücksichtigung der Profil-Korrelation abgeändert, sodass die Zuweisung eines Nutzers nicht mehr zum nächstgelegenen Zentrum erfolgt. Bei Beachtung der maximalen Entfernung ergibt sich die Zielfunktion gemäß Gleichung 2.5.

$$D_{center}(K) = \min \max_{v_i=1,\ldots,n} d(v_i, k_u) \tag{2.5}$$

Bei Beachtung der durchschnittlichen Entfernung ergibt sich die Zielfunktion gemäß Gleichung 2.6.

$$D_{median}(K) = \min \sum_{v_i=1}^{n} d(v_i, k_u) \tag{2.6}$$

2.4 Voraussetzungen und Annahmen

Bei der betrachteten Problemstellung werden folgende Voraussetzungen als gegeben angesehen und Annahmen für die Anwendbarkeit getroffen. Für das jeweilige betrachtete Szenario sind die Standorte der Nutzer sowie deren Interessen-Profile gegeben. Darüber hinaus liegt die Information über die Infrastruktur vor, auf der das System aufsetzt. Wenn eine explizite Infrastruktur vorgegeben ist, dann genügt diese mindestens der Definition zusammenhängend. Die Knoten sind demzufolge paarweise durch eine Verbindung miteinander verbunden. Während der Optimierung wird die Infrastruktur als statisch betrachtet, da eine Veränderung eine andere Aufgabenstellung beschreibt. Auf Grund der NP-schwere des Problems sind bei einer geänderten Infrastruktur die Berechnungen für eine Lösung zu wiederholen. Die durch einen Betreiber bereitgestellte Infrastruktur bietet ausreichend Leistung in Anbetracht der Kapazität und Bandbreite, um die Verbindung zwischen Nutzer und Zentrum zu gewährleisten. Da die Infrastruktur von einer anderen Instanz verwaltet wird, bietet diese somit einen Service für das Transportsystem an und gewährleistet die dafür notwendige Zuverlässigkeit und Qualität. In dieser Arbeit erfolgt vorrangig eine Betrachtung der Planungsphase des Systems für die Transportkette. Daher wird eine hohe Verfügbarkeit der einzelnen Systembestandteile vorausgesetzt. Das System wird für eine lange Laufzeit ausgelegt. Bei der Bereitstellung eines Dienstes durch ein Zentrum muss dieses den Dienst zuvor in seinem Speicher ablegen. Es erfolgt keine direkte Weiterreichung ohne vorherige Aufnahme im Zentrum. Gemäß dem Ziel, das System hinsichtlich der Zufriedenstellung der Bedürfnisse von Nutzern auszurichten, erfolgt die Speicherung der Dienste in einem Zentrum nutzergetrieben und bedarfsorientiert.

3 Anforderungsanalyse

Aus den Anwendungsfällen und den wissenschaftlichen Fragestellungen ergeben sich mehrere Anforderungen hinsichtlich einer Lösung und der durchzuführenden Analysen. Die Anforderungen lassen sich in funktional und nicht-funktional unterteilen. Zudem sind systemübergreifende Anforderungen aufgeführt. Neben den zu erfüllenden Leistungen sind zur Abgrenzung des Forschungsbereiches die Randbedingungen als Bewertungskriterien zusammengefasst.

3.1 Nicht-Funktionale Anforderungen

Gemäß der Problemstellung besteht das Ziel in der Entwicklung eines generischen Modells für zweistufige Transportketten. Die Bezeichnung generisch bezieht sich hierbei auf die flexible Anwendbarkeit des Modellierungsansatzes für mehrere Bereiche. Mit dem generischen Modell sind Prozesse nachzubilden und Aufgabenstellungen zu lösen, welche trotz unterschiedlicher Nebenbedingungen nach einer gemeinsamen Funktionsweise ablaufen. Ohne Beschränkung der weiteren Nutzbarkeit ist das Modell hinsichtlich der betrachteten Anwendungsdomänen FLP und CDN auszurichten. Um den domänenspezifischen Eigenschaften Rechnung zu tragen, muss sich das Modell über Parameter an mehrere Nebenbedingungen anpassen lassen. Für ein spezifisches Szenario sind die verschiedenen Eingangsdaten gemäß einer einheitlichen Schnittstelle zu verarbeiten. Diese Informationen werden für die Erzeugung des Modells auf der Grundlage eines generischen Ansatzes verwendet.

3.2 Funktionale Anforderungen

Der Prozess der Modellerzeugung ist hinreichend zu spezifizieren, sodass dieser zur Aufwandsminimierung bei neuen Aufgabenstellungen eine Automatisierung zulässt. Zur Planung von Zentren muss das Modell die Optimierung der Standorte hinsichtlich eines spezifischen Kriteriums als auch eine multikriterielle Optimierung ermöglichen. Das allgemeine Ziel ist dabei die Verbesserung der QoS für den Nutzer bei minimiertem Einsatz beschränkter Ressourcen.

Unter Beachtung der bei dieser Problemstellung auftretenden Fragestellungen sind speziell die vier wesentlichen Optimierungsaspekte zu analysieren. Dazu hat eine Untersuchung hinsichtlich der Anzahl von Zentren für ein Szenario zu erfolgen. Die Standorte sind unter Beachtung der parametrierten Nebenbedingungen zu optimieren. Dies beinhaltet die Zuweisung der Nutzer zu einem primären Zentrum, zu dem ein

Nutzer seine Anfragen stellt. Die Beziehung zwischen Zentrum und Nutzer ist hierbei als bidirektional anzusehen, da ein wechselseitiger Dienstfluss erfolgt. Die dazu notwendigen Prozesse, insbesondere die Optimierung und die Anwendbarkeit der Algorithmen bei unterschiedlichen Nebenbedingungen, sind entsprechend zu spezifizieren. Auf Grund der Größe realer Aufgabenstellungen und der NP-schwere des Problems sind die einzelnen Prozessschritte flexibel anpassbar auszulegen. Die Komplexität der Problemstellung, insbesondere die Platzierung von Zentren mit der Zuweisung von Nutzern, erfordert eine automatische Optimierung. Die verschiedenen Möglichkeiten der Zuweisung und der daraus resultierenden Auswirkungen auf das Systemverhalten erfordern eine simulative Analysemöglichkeit der Prozessabläufe. Für die Simulation ist eine Vorabspezifizierung der Kapazität der Zentren sowie der Verwaltungsstrategie der Speicher erforderlich. Neben der generischen Modellierung ist für diese Arbeit die Berücksichtigung der spezifischen Nutzereigenschaften und -profile von maßgeblicher Bedeutung. Diese Aspekte sind zur Leistungssteigerung der Gesamtsystemleistung mit in die Optimierung einzubeziehen. Für die Analyse von operativen Effekten und Auswirkungen zur Laufzeit müssen sich in der Simulation die spezifischen Modelleigenschaften auswirken. Insbesondere ist hierbei das Speicherverhalten bei den Zentren zu beachten. Dies erfordert die funktionale Modellierung der Speicherungsprozesse eines Zentrums. Bei der Anfrage eines Nutzer nach einem Dienst ergeben sich mehrere Möglichkeiten in der Bereitstellung durch ein Zentrum. Diese sind entsprechend in der Simulation zu realisieren.

3.3 Systemübergreifende Anforderungen

Die Eigenschaften des geforderten Modells werden in fall-spezifische, domänen-abhängige und domänen-unabhängige Bestandteile unterschieden. Durch diese Abstraktionsstufen lässt sich eine hohe Wiederverwendbarkeit der einzelnen Bestandteile und eine bessere Handhabbarkeit der Komplexität erreichen. Jedes Szenario besitzt mehrere Eingangsdaten in unterschiedlicher Ausprägung, welche konkret für den einzelnen Fall spezifisch sind. Dazu gehören unter anderem die dedizierten Positionen der Nutzer und deren Interessen-Profile. Diese Informationen sind mit den Nebenbedingungen der spezifischen Domäne zu ergänzen. Dies ist über Parameter zu steuern, welche in Abhängigkeit des Anwendungsgebietes parametriert werden. Die Menge der Parameter sowie deren Bereiche sind dafür vorab so zu spezifizieren, dass diese den generischen Forderungen des Modells genügen. Mittels dieser Eingangsdaten ist gemäß des geforderten Ansatzes ein Modell zu erstellen, welches domänen-unabhängig repräsentiert wird. Das generische Modell ist über eine einheitliche Schnittstelle zu optimieren. Dazu sind die verschiedenen Lösungsansätze entsprechend der spezifizierten Schnittstelle an das Modell anzuschließen. Die Abbildung 3.1 stellt die Unterteilung in die einzelnen Bereiche dar und zeigt deren Zusammenhang.

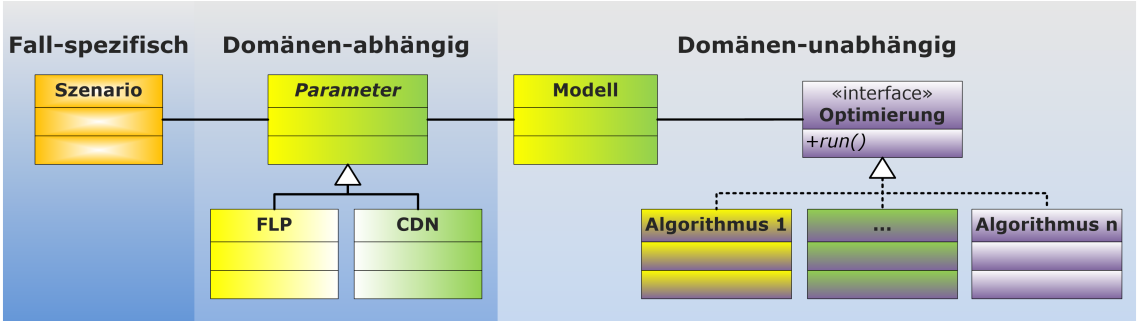


Abbildung 3.1: Struktur des generischen Modellierungsansatzes zur Optimierung von Problemstellungen aus den Bereichen FLP und CDN sowie weiterer Anwendungsgebiete von zweistufigen Transportketten

3.4 Bewertungskriterien

Die folgende Auflistung fasst die beschriebenen Anforderungen zusammen und dient einer vereinfachten Kategorisierung von bestehenden Arbeiten. Diese Bewertungskriterien sind von der betrachteten Problemstellung im Kapitel 2 abgeleitet und entsprechend der Zielstellung des generischen Modells im Kapitel 1.4 aufgearbeitet.

- **Anwendungsgebiet**
  Ein betrachtetes Modell ist für ein spezifisches Anwendungsgebiet ausgelegt und beschreibt die Erweiterbarkeit und Anpassbarkeit für weitere Bereiche.
- **Anzahl der Zentren**
  Die zu platzierende Anzahl von Zentren lässt sich variieren und eine Analyse dessen Einfluss auf das Gesamtsystem ermöglicht.
- **Optimierte Platzierung der Zentren**
  Unter Beachtung von verschiedenen Nebenbedingungen erfolgt eine Selektion der Standorte und optimierte Platzierung von Zentren.
- **Wechselseitige Abhängigkeit**
  Der Standort und Typ eines Zentrums lässt sich unter Berücksichtigung der Entfernung zu den anderen Zentren über Parameter beeinflussen und planen.
- **Differenzierung von Verbindungspfaden und Nutzern**
  Es wird zwischen verschiedenen Verbindungspfaden der Infrastruktur unterschieden und die Wichtigkeit einzelner Nutzer für den Systembetreiber lässt sich differenzieren.
- **Last auf die Infrastruktur**
  Das Modell berücksichtigt vorgegebene Infrastrukturen und bezieht diese in die Optimierung mit ein, sodass Rückschlüsse auf die voraussichtliche Auslastung möglich sind.
- **Zuweisung der Nutzer**
  Ein Nutzer wird spezifisch einem Zentrum anhand mehrerer Kriterien und Leistungsparameter zugewiesen.

- **Nutzerprofile**
  Die Nutzer besitzen ein individuelles Interessen-Profil, welches im Modell abgebildet ist und während der optimierten Lösungsberechnung berücksichtigt wird.
- **Bereitstellung des Dienstes**
  Das Modell überwacht, ob ein Zentrum einen angeforderten Dienst überhaupt bereitstellen kann.
- **Speicherkapazität der Zentren**
  Die Speicherkapazität der Zentren wird im Modell berücksichtigt und das System ermöglicht eine Variation der Kapazität.
- **Simulation**
  Es erfolgt eine Simulation des Transportsystems und der berechneten Lösung zur Ermittlung verschiedener Leistungsparameter.

4 Stand der Wissenschaft

In der Literatur gibt es für einzelne Teilbereiche der beschriebenen Problemstellung mehrere Lösungsansätze. Diese werden im Folgenden entsprechend deren Anwendungsgebiet unterschieden und erläutert. Die Einteilung erfolgt anhand der Anwendungsgebiete in Systeme für die Bereiche Logistik und Rechnernetze. Im Bereich der Logistik gibt es neben der Organisation als Transportkette noch eine weitere Art, welche als Nabe und Speiche Systeme bezeichnet werden. Bei dieser Form erfolgt die Interaktion nicht ausschließlich vom Zentrum zum Nutzer, sondern verläuft zwischen den Nutzern selbst. Die Zentren nehmen dabei die Rolle der Vermittler ein, sodass ein Transport stets über das jeweils zugewiesene Zentrum eines Nutzers erfolgt. Somit ist die Verbindung zwischen den Zentren zu berücksichtigen. Darüber hinaus erfolgt eine Betrachtung der wissenschaftlichen Arbeiten zur Modellierung von Nutzerverhalten. Abschließend erfolgt eine Betrachtung der ganzheitlichen Ansätze für das Problem, bevor die wesentlichen Arbeiten tabellarisch in einer Übersicht zusammengefasst werden. Zur Lösungsberechnung der Problemstellung sind in einem weiteren Abschnitt die verschiedenen Algorithmen und Optimierungsansätze erläutert.

4.1 Logistik - Facility Location Problem

Im diesem Abschnitt werden wissenschaftliche Arbeiten im Bereich des FLP einzeln vorgestellt und eine Abgrenzung zu der vorliegenden Arbeit beschrieben.

Zur Minimierung der maximalen Entfernung beim FLP erfolgt in der Veröffentlichung [71] eine Platzierung der Zentren mittels Algorithmen, welche eine garantierte Lösungsgüte des fünf- bzw. sechs-fachen des Optimums besitzen. Es werden zwei spezifische Problemstellungen der Knotenplatzierung betrachtet, wobei ein Zentrum eine maximale Anzahl an Nutzern bedienen darf. Dabei ist es gestattet, dass mehrere Zentren an dem gleichen Standort platziert werden. Die Algorithmen nutzen die Baumstruktur der Szenarien und sind somit für das in dieser Arbeit vorgestellte Problem nicht anwendbar.

In den Studien [50, 72] werden Modelle für verschiedene Möglichkeiten der Platzierung von Zentren aufgezeigt. Die Übersicht teilt die Lösungsansätze in die vier Kategorien planare Modelle, lagerbasierte Modelle, netzbasierte Modelle und diskrete Modelle ein. Die unendliche Anzahl von Standort-Möglichkeiten bei planaren Modellen wird schrittweise bis hin zur direkten Vorgabe von einzelnen Positionen bei diskreten Modellen eingeschränkt. Die Arbeit zieht mit den zahlreichen Quellen ein Resümee über den derzeitigen Stand der Problemstellung. Hierbei werden keine direkten Lösungsvorschläge gegeben, sondern die Probleme ausführlich definiert. Vor allem wird nicht auf die Eigenschaften des Zentrums näher eingegangen. Dennoch zeigt die Arbeit anhand der diskreten Modelle

die fundamentale Komplexität der Probleme auf. Angenommen es gibt 2^n Möglichkeiten der Platzierung von k Zentren. Da es sich um ein NP-Problem handelt, müssen zur Identifizierung der optimalen Lösung alle Lösungsmöglichkeiten evaluiert werden. Wenn mit $n = 50$ verschiedene Entscheidungen zu treffen sind und jede sich innerhalb einer Nanosekunde berechnen lässt, dann dauert die vollständige Berechnung mehr als 13 Tage. Bei einer realistischeren Annahme von einer Million Auswertungen pro Sekunde dauert die Berechnung über 35 Jahre.

Eine komplexe Logistik-Kette wird in der Arbeit [73] simuliert. Hierbei verlangen die Konsumenten des Systems nach einem einzelnen Produkt, welches durch die mehrstufige Transportkette bedient wird. Die betrachtete Struktur ist vergleichbar mit einem gerichteten Netz. In der Simulation wird der Transport und der Warenfluss im Detail abgebildet. Der Fokus der Optimierung liegt auf der Reduktion der Simulationskomplexität zur Verringerung der Simulationszeit und zur Steigerung der Simulationseffizienz. Im Vergleich zu anderen Ansätzen hat die räumliche Beziehung der einzelnen Teilnehmer der Transportkette keine Relevanz. Weiterhin ist das Modell nur für eine Produktart ausgelegt, wodurch ein Einfluss der komplexen Lagerhaltung nicht gegeben ist.

Der Autor Daskin beschreibt in seinem Buch [74] umfassende Modelle zur Standortplanung im Anwendungsbereich Logistik. Hierbei werden vorrangig Modelle auf Basis der linearen Programmierung skizziert, aber auch gängige, heuristische Verfahren erläutert. Insbesondere erfolgt eine Betrachtung von Mehrziel-Problemen. Diese beziehen sich auf die Nebenbedingungen bei der Platzierung von Zentren in Bezug auf die Infrastruktur und geänderte Prozesse beim Transport. Eine Betrachtung des Nutzerverhaltens erfolgt dabei nicht. Auch eine Anwendbarkeit der Modelle bei stochastischen Einflüssen oder die Übertragbarkeit auf weitere Bereiche ist nicht gegeben.

Die Arbeit [75] zur Standortplanung beschreibt ein mathematisches Modell mittels *Linear Programming (LP)* zur Steigerung der Dienstgüte. Hierbei wird deutlich gemacht, dass es mit zunehmender Entfernung eines zu versorgenden Standortes vom Zentrum zu einem Verlust an Dienstqualität kommt. Das Modell ist spezifisch für den Anwendungszweck konstruiert. Dies gilt ebenfalls für das spezifische Gravitation p-Modell [76], bei dem Kunden in Abhängigkeit der Entfernung und der Attraktivität von Einrichtungen angezogen werden. Der Ansatz zur Lokalisierung der zentralen Einrichtungen wurde für mehrere Szenarien in Schweden angewendet, wobei der Fokus auf der Standortbestimmung liegt und angrenzende Effekte außer Acht gelassen sind. Die Ergebnisse entsprechen denen des gängigen p-median Modells, wodurch im Folgenden das allgemeinere p-median Modell zu beachten ist.

In der Übersicht [77] werden Modelle zur Planung von Verteilzentren unter Berücksichtigung von Unsicherheiten der Eingangsdaten betrachtet. Im Laufe der Zeit ändern sich die Anforderungen und es gibt Schwankungen im Bedarf, sodass bei der Planung diese Unsicherheiten entsprechend mit einbezogen werden. Dafür bieten die beschriebenen Modelle einen Ansatz, um dennoch eine Planung hinsichtlich eines langen Zeithorizontes durchzuführen. Der Fokus liegt dabei auf stochastischen und robusten FLP Modellen, wobei keine Analyse mittels Simulation verwendet wird und der Ansatz auf FLP eingegrenzt ist.

In den Büchern [78, 79] wird das k-center Problem sowie die gewichtete Variante grundlegend beschrieben. Zudem enthalten diese den Beweis, dass kein Algorithmus mit einem Polynomialzeit-Verhalten eine garantierte Leistung besser als zwei-approximiert erreichen kann, solange P $\neq$ NP gilt. Darüber hinaus wird mit dem Greedy-Algorithmus ein solcher darin beschrieben. Für das FLP in der Ausprägung k-center und k-median sind LP-Relaxation angegeben, welche auch zum Vergleich der statischen Optimierung herangezogen werden. Weiterhin stellt die Arbeit [80] in diesem Zusammenhang verschiedene heuristische Verfahren gegenüber. Dabei wird die Leistung des Greedy-Ansatzes für geringe Anzahlen von Zentren im Vergleich zu anderen getesteten Verfahren hervorgehoben.

In der Dissertation [81] werden Modelle für Flow-Interception im Bereich Geoinformationssysteme beschrieben. Hierbei unterbrechen Verbraucher ihre geplante Reise, um ein Produkt oder eine Dienstleistung zu erhalten. Dies geschieht jedoch nur an bestimmten Stellen in der Nähe von der Route, sodass sich der Umweg in Grenzen hält. Bei diesen werden Standorte für Dienste gesucht, um möglichst vielen Kunden eine Unterbrechung gemäß deren Toleranzverhalten zu ermöglichen. Abgesehen vom Modellierungsansatz zur Platzierung erfolgt hinsichtlich der betrachteten Problemstellung in der vorliegenden Arbeit keine nähere Analyse.

Die Veröffentlichung [82] beschreibt ein Modell zur Optimierung von Logistiknetzen zur Verbesserung des Servicelevels bei gleichzeitiger Reduktion der Kosten. Dazu erfolgt eine Betrachtung mittels integrierter Entscheidungsunterstützung zur Platzierung der Zentren, der Bestandssteuerung und der Transportroute. Das Modell an sich ist über Gleichungen gemäß *Integer Linear Programming* definiert. Über eine Partikel Schwarm Optimierung wird die Standortbestimmung vorgenommen. Der Ansatz ist spezifisch für ein Szenario in der Logistik mit konkaver Kostenfunktion, sodass dieser sich nicht auf die in der vorliegenden Arbeit fokussierte Problemstellung übertragen lässt. Dennoch wird die Notwendigkeit einer integrierten Lösungsfindung ersichtlich.

In der Arbeit [83] ist ein Modell einer Lieferkette beschrieben. Entwurfs- und Planungsprobleme sind in einem einzigen kombinierten System abgebildet. Dieses besteht aus den vier Teilen Design des Liefernetzes, Planung für Produktion sowie Verteilung vom Lieferanten zum Kunden, Modellmanagement und Datenmanagement. Die Probleme werden dabei unabhängig voneinander sequentiell gelöst. Dabei kommen eine Lagrange-Heuristik und ein genetischer Algorithmus zum Einsatz. Das mathematische Modell betrachtet die Kosten des Materials und ist spezifisch für den Warentransport in der Logistik.

Das Buch [84] gibt einen umfassenden Einblick in die Planung von Logistiknetzen. Über die verschiedenen Planungstätigkeiten hinweg werden die typischen Prozesse und Effekte beschrieben. Schwerpunkt ist dabei die Versorgungssicherheit durch Lagerhaltung und das operative Supply Chain Management. Zudem wird unter anderem ein Überblick zur Standortbestimmung in der Ebene und auf Netzen gegeben, jedoch keine spezifischen Strategien vorgestellt. Weiterhin fehlen übergreifende Ansätze, um eine integrierte Lösung zu erreichen.

Das Lehrbuch [85] bietet einen grundlegenden Ansatz zur Modellierung von Logistiknetzen mittels Graphentheorie. Dazu werden die einzelnen Bestandteile entsprechend abstrahiert und in einem Modell umgesetzt. Neben der allgemeinen Beschreibung von Transportketten erfolgt eine spezifische Betrachtung von Modellen zur Standortwahl. Als Lösung wird dabei das Simplex-Verfahren unter Nutzung von Kantengewichten beschrieben, welches in der vorliegenden Arbeit als Referenz dient. Darüber hinaus erfolgt keine Beachtung von Nutzerprofilen oder der dynamischen Bereitstellung von Waren.

4.2 Logistik - Nabe-Speiche-Systeme

Bei Nabe-Speiche-Systemen (HaS, engl. Hub and Spoke) werden Güter zwischen zwei Partnern A und B nicht direkt untereinander ausgetauscht. Der Transport von Gütern erfolgt über zentrale Knotenpunkte. Eines der größten und weit verbreitetsten Beispielsysteme ist das der Post. Hierbei werden Briefe und Pakete von den lokalen Annahmestellen zu einem Zwischenlager transportiert indem Fahrzeuge diese im lokalen Bereich abholen. Bei HaS ist jeder lokale Netzknoten einem oder mehreren Zwischenlagern zugeordnet. Im Zwischenlager wird anhand der Zieladressen der weitere Transportweg bestimmt. Ziel ist der direkte Weitertransport jeder Ware zu dem Zwischenlager, in dessen Einzugsbereich die Zieladresse liegt. Nachdem die Güter an den entsprechenden Zwischenlagern angekommen sind, werden diese an die entsprechende, lokale Zieladresse ausgeliefert. Auch wenn der direkte Transportweg zwischen zwei Systemknoten kürzer und schneller wäre, erfolgt bei HaS der Transport stets über die entsprechenden Zwischenlager. Das Problem besteht in der Standortplanung der Zwischenlager, wobei dies in dem speziellen Fall als Hub Location Problem bezeichnet wird [86]. Dies beinhaltet die entsprechende Zuweisung der Systemknoten sowie in dem lösen des Traveling-Salesmans Problems [87] bei der Transportroute eines Zwischenlagers und den zugewiesenen Knotenpunkten. Dazu sind in diesem Abschnitt wissenschaftliche Arbeiten erläutert und Bezüge zu der betrachteten Problemstellung aufgezeigt.

Die Arbeit von Mayer [88] untersucht HaS mittels quantitativer Methoden zur Minimierung von Transportkosten. Hierbei bezieht sich der Ansatz vor allem auf den Transport von Paketen im Luftverkehr. Die Optimierung des Systems ist auf die Standortbestimmung für die Hubs und die Planung von Transportverbindungen ausgelegt. Die dazu eingesetzten Verfahren umfassen heuristische Ansätze sowie exakte Lösungsverfahren. Dazu gehören z. B. Standort-Austausch-Verfahren, Tabu-Search und Evolutionäre Algorithmen sowie Branch-and-Bound und Floyd-Washall-Verfahren. Darüber hinaus sind Entscheidungsunterstützende Instrumente beschrieben, deren Leistung verglichen wird. Allerdings handelt es sich bei dieser Betrachtung um ein statisches Modell, welches weder spezifische Nutzerinteressen berücksichtigt noch die Zuordnung genauer analysiert. Eine flexible Anpassung des Modells auf weitere Anwendungsdomänen ist auf Grund des spezifischen Fokus nur eingeschränkt möglich.

Der öffentliche Personen-Nahverkehr wird bei dem Ansatz von Owsinski [89] als zu minimierendes Hub Problem dargestellt. Das auf Graphen basierende Modell

berücksichtigt Prioritäten von Knoten und Gewichte der Verbindungen. Im Gegensatz zu den anderen Modellen wird dadurch ein höherer Detailgrad erreicht. Neben dem Vergleich mit dem *Dominating Set* und dem *Minimum Set Covering* Problem werden Grenzen der Lösungsberechnung aufgezeigt. Mittels Greedy Verfahren und Evolutionären Algorithmen werden für ein reales Szenario Lösungen optimiert. Dieses Modell ist spezifisch für das Szenario erstellt, sodass es sich nicht auf die in dieser Arbeit betrachtete Problemstellung anwenden lässt.

4.3 Rechnernetze - Content Delivery Networks

Dieser Abschnitt beschreibt einzelne wissenschaftliche Arbeiten im Bereich von CDN und zeigt die Unterschiede zu der vorliegenden Arbeit auf.

Der Modellierungsansatz [90] konzentriert sich auf *Content Centric Networking* (CCN). Hierbei werden Speicher für einzelne Pakete an allen routenden Netzkomponenten eingefügt und nach dem *Least Recently Used* (LRU) Prinzip bzw. dem eigenen Verfahren *Latency-Aware Caching* (LAC) [91] verwaltet. Das Analysemodell nutzt zur Abbildung einen zweistufigen Markov modulierten Ratenprozess mit einer nach der Zipf-Funktion verteilten Popularität von Datenpaketen. Die betrachteten Netzstrukturen sind lineare Ketten und Bauminfrastrukturen, sodass der Ansatz nicht ohne weiteres auf beliebige Infrastrukturen übertragbar ist. Das Ergebnis ist eine statische Analyse der Transferleistung in Abhängigkeit von der Popularität der Dateninhalte in Relation zur Speichergröße. Dieser Ansatz zeigt, dass die Transferzeit von der Länge der Netzpfade abhängt und die Zipf-Funktion sich zur Simulation von Nutzeranfragen eignet, nimmt jedoch keine dedizierte Platzierung von Zwischenspeichern vor.

Die Arbeit [56] untersucht die optimierte Platzierung von Speichern in Rechnernetzen. Das Ziel besteht in der Minimierung der Netzlast und der Zugriffszeit durch den Einsatz von einer spezifischen Anzahl von Zwischenspeichern. Der Fokus liegt dabei auf den Netzstrukturen Linie und Ring sowie in der theoretischen Betrachtung von Baumstrukturen. Die Platzierung der Zwischenspeicher erfolgt vorrangig nach dem Prinzip des Greedy Algorithmus. In dem Modell senden die Clients deren Anfragen stets an den nächstgelegenen Server, welcher für die Bereitstellung der Daten zuständig ist. Es wird gezeigt, dass sich die optimale Position der Speicher über die Zeit ändert. Der Ansatz ist spezifisch für CDN konzipiert.

Daran anschließend betrachten die Ansätze [92–94] die Platzierung von Inhalten auf vordefinierten Servern. Dabei erfolgt eine Optimierung der Antwortzeit von Anfragen und des Lastausgleichs zwischen den Servern mittels Greedy Algorithmus. Die zu Grunde gelegte Infrastruktur ist dabei allgemeiner gehalten als bei [56]. Über ein mathematisches Modell wird im Ansatz [95] dazu eine statische Analyse gegeben. In Relation zu der Problemstellung der vorliegenden Arbeit sind diese spezifisch für den Anwendungsbereich und umfassen nicht alle der hier betrachteten Aspekte.

Gemäß der Arbeit [96] befindet sich die optimale Position von Servern innerhalb des Backbone Netzes anstelle den Übergängen zu anderen Netzen. Die Gesamtreduktion der Netzlast ist bei Zwischenspeicherung von Daten an Kernknoten innerhalb des

Netzes höher als bei externen Knoten. Diese wurde anhand von FTP-Datenverkehr gezeigt. Demgegenüber hat gemäß der Analyse [97] eine Platzierung der Speicher an Netzübergängen zu erfolgen. Der Nachweis erfolgt mittels einer kooperativen Cache-Management-Strategie über ein *Linear Programming* Ansatz. Daher wird in der vorliegenden Arbeit die Platzierung als Optimierungskriterium aufgefasst und für den Anwendungsbereich CDN die Analysen anhand der Infrastruktur von Backbone Netzen durchgeführt.

In der Dissertation [98] werden auf Basis von Traces Techniken entwickelt, um die Leistung von verteilten Informationssystemen zu verbessern. Der Fokus liegt dabei auf der Entwicklung von Datenbanken für Traces beim Client. Dazu erfolgt eine breitflächige Analyse des Anwendungsgebietes im Bereich des World-Wide Web. Auf der Client-Seite werden dazu Vorhersage-Techniken angewendet, damit die Daten bereits vor der eigentlichen Anfrage zur Verfügung stehen und somit sich die Antwortzeiten verringern. Auf der Server-Seite erfolgt eine optimierte Platzierung der Dateninhalte zur Reduktion der Bandbreitenauslastung und dem Lastausgleich unter Servern. Im Bezug zur vorliegenden Arbeit ist bei der Entwicklung des generischen Ansatzes eine Erweiterbarkeit hinsichtlich solcher Ansätze zu achten.

Für CDN mit baumartigen Netztopologien werden in der Arbeit [99] verschiedene Strategien zur Platzierung von Servern beschrieben. Das Ziel des Ansatzes besteht in der Minimierung des erzeugten Datenverkehrs unter Beachtung einer maximalen und durchschnittlichen Kommunikationsentfernung. Es wird dabei die Unterscheidung zwischen der diskreten Platzierung an einem Knoten in leitungsgebundenen Netzen und der kontinuierlichen Platzierung an einer Verbindung bei drahtlosen Netzen vorgenommen. Der strikte Fokus auf Infrastrukturen hinsichtlich Bäumen lässt sich nicht verallgemeinern.

Im Bereich von SDN untersucht die Arbeit [100] die Platzierung von SDN-Controllern zum Management des Netzes mit dem Ziel einer Reduktion der zu erwartenden Last. Dies geschieht anhand der maximalen Entfernung entsprechend dem k-center Problem mittels einem Linear Programming Modell. Hierbei findet keine multikriterielle Optimierung statt.

Im Bereich von CCN beschreiben die Arbeiten [101–103] die Vorteile von Cache-Infrastrukturen, insbesondere den schnelleren Datenzugriff. Bei diesem inhaltsorientierten Ansatz werden die übertragenen Daten adressiert und routbar gestaltet. Anhand von Szenarien mit Binär- und Ternärbaumtopologie erfolgen Analysen zu dessen Leistungsfähigkeit. Dabei wird zwischen statischer und dynamischer Speichereinteilung unterschieden und die Ersetzungsstrategien LRU und *Least Frequently Used* (LFU) eingesetzt. Gemäß einem Ansatz für Datenbanken [104] liefert LRU dabei nahezu optimales Management. Das Ziel besteht in der Verbesserung der QoE für den Endnutzer. Jedoch erfolgt keine optimierte Platzierung.

Die Messergebnisse des Harvest Projektes [105] zeigen die Skalierbarkeit von hierarchischen Speicherstrukturen, insbesondere für Web- und FTP-Daten. Das Projekt bezieht sich auf eine Speicherstruktur gemäß der Anordnung eines Baumes und verdeutlicht die Vorteile von Zwischenspeichern im Vergleich zu den Mehrkosten. Die Analyse stellt die Verringerung der notwendigen Bandbreite sowie der durchschnittlichen Zugriffszeit bei der Nutzung von Proxy-Servern heraus, bietet aber keinen Modellie-

rungsansatz. Das Verhältnis des Aufwandes eines Cache-Miss zu einem Cache-Hit bei einem Zwischenspeicher wird hierbei mit 10:1 bei der Leistung von Internet Informationssystemen angegeben.

Die Arbeit [106] beschreibt eine Architektur zur Anordnung der Server in einem Ring mit dem Ziel einer effizienten Informationsverbreitung. In Verbindung mit einem P2P Netz wird ein Vergleich mit einem zentralisierten Ansatz hervorgehoben. In den Analysen variiert die Speicherkapazität der Server von 5 % bis 45 % des gesamten Inhalts. Zur Simulation des Nutzerverhaltens wird das Modell der Mandelbrot-Zipf-Verteilung mit $\beta = 0.74$ eingesetzt [107, 108]. Der Leistungsvergleich [109, 110] zur Inhaltsplatzierung zeigt für dieses Szenario die Komplexität des Problems. Im Vergleich zur vorliegenden Arbeit liegt der Fokus auf der Organisation der Server im operativen Betrieb anstelle der Vorabplanung.

Das IDMaps Projekt [111] beschäftigt sich mit der Erstellung von Entfernungskarten des Internets. Dabei werden graphentheoretische Ansätze und ad hoc Heuristiken zur Abbildung mittels sogenannter Tracers als Address Prefixes eingesetzt. Für die Platzierung der Tracers ergeben sich ähnliche Fragestellungen wie bei der direkten Platzierung von Servern in einem Netz. Zur Evaluation erfolgt die Identifizierung von zentralen Knotenpunkten auf den ermittelten Karten im Vergleich zu den vollständig bekannten Topologien [112]. Die Analyse zeigt, dass für eine effiziente Platzierung die vollständige Infrastruktur nicht bekannt sein muss. Darüber hinaus werden keine weiteren Kriterien betrachtet.

Die Bücher [113, 114] geben einen umfassenden Überblick zu CDN und einen technischen Einblick in deren Funktionsweise, wobei die Planung außer acht gelassen ist. Neben einer Taxonomie wird die Leistungsfähigkeit hinsichtlich der Art und Weise von der Zwischenspeicherung von Inhalten als auch das Anfragen-Management betrachtet. Dabei erfolgt eine allgemeine Bewertung zur statischen und dynamischen Speicherung von Inhalten. Weiterhin sind die zukünftigen Community basierten CDN erläutert. In beiden Werken wird auf die Platzierung von den Cache-Servern nahe des Nutzers hingewiesen, um die Zugriffszeit zu minimieren. Bei [114] wird explizit auf die Platzierung von den Servern eingegangen. Weiterhin wird motiviert, dass die Platzierung nicht ausschließlich anhand der Zugriffszeiten und der Verbindungsstrecken auszurichten ist, sondern auch die Interessen der Nutzer mit einzubeziehen sind. Allerdings werden dafür keine Lösungsansätze gegeben.

Einer der größten Vertreter von CDN ist Akamai [115], welcher bis zu 20 % des weltweiten Datenverkehrs von Webseiten erbringt. Das Paper gibt einen Überblick zur Plattform und beschreibt die Prozesse, die für den generischen Modellierungsansatz umzusetzen sind. Es stellt die Bedeutung von kurzen Netzpfaden heraus und erläutert die Notwendigkeit von einer hohen Servicequalität. Gemäß einer Untersuchung warten 40 % der Besucher einer Website nicht länger als drei Sekunden zum Laden, bevor der Aufruf abgebrochen wird [116]. Zur Identifizierung von kurzen Netzpfaden im Overlaynetz setzt Akamai *Linear Programming* Ansätze ein [117]. Auf Grund des spezifischen Anwendungsbereiches und des Fokus auf Netze ist die Arbeit nicht als generischer Ansatz einsetzbar. Die Kooperation von Akamai als CDN Anbieter mit der Deutschen-Telekom als ISP motiviert die Notwendigkeit multikriterieller Lösungen [118].

Die Dissertation [119] beschreibt ein Modell zur Platzierung von Inhalten auf Servern eines CDN. Dazu wird das entwickelte *Linear Programming* Modell gegenüber den Optimierungsverfahren *Simulated Annealing* und *Variable Neighborhood Search* gestellt. Dagegen fokussiert die Dissertation [120] auf die gleichmäßige Auslastung der Ressourcen zur Verteilung von Daten und geht dabei auf das Nash-Gleichgewicht ein. Beide Ansätze, wie auch das Verfahren [121], stellen lediglich eine spezifische Lösung für ein statisches Problem dar. Für einen dynamischeren Ansatz bietet die Arbeit [122] einen neuen Optimierungsalgorithmus zur Replizierung und Platzierung von Inhalten, wobei die Platzierung und Funktionalität der Server nicht analysiert werden. Die Evaluation erfolgt dabei anhand der Auswirkungen von sogenannten Flash Crowds.

Einer Übersicht zur Funktionsweise von CDN geben die Arbeiten [123, 124]. Neben gängigen Praktiken vermittelt diese eine Idee zur Balancierung von Kosten eines CDN-Anbieters und Content-Providers bei gleichzeitiger Beachtung der QoS für Benutzer von Webdiensten. Allerdings wird hierbei die Rolle des ISPs nicht betrachtet und es fehlt an detaillierten Vorschlägen zur Erreichung der Optimierungsziele. Dennoch unterstreicht deren Untersuchung die Bedeutung einer angepassten Platzierung von Servern mit kurzen Entfernungen zu Nutzern und dem Management der gespeicherten Daten. Insbesondere wird das dynamische Zwischenspeichern von Inhalten als Optimierungspotential hervorgehoben, um auf die individuellen Präferenzen von den Nutzern einzugehen.

Die Dissertation [125] beschäftigt sich mit dem Einfluss von Diensten auf die Infrastruktur und bietet Lösungsansätze zur Berechnung optimierter Pfade durch das Netz. Hierbei erfolgt die Ermittlung einer Overlay Infrastruktur zur gesamtheitlichen Verbesserung der Kommunikation in einem Netz. In Bezug zu der vorliegenden Arbeit ist somit davon auszugehen, dass der Betreiber der Infrastruktur sich nach den Anforderungen der Nutzer richtet sowie seinen Infrastruktur-Dienst entsprechend ausrichtet und bereit stellt. Trotz des spezifischen Schwerpunktes lassen sich solche Ergebnisse bei einem generischen Ansatz weiter nutzen. Dazu hat der neue Ansatz Schnittstellen anzubieten, worüber sich bisherige Erkenntnisse weiterverwenden und in neue Anwendungsbereiche übertragen lassen.

Demgegenüber erfolgt in der Arbeit [126] eine Simulation von CDN und die Entwicklung einer Architektur für ein verteiltes CDN. Hierbei wird für den Lastausgleich zwischen den Servern eine Strategie beschrieben und über ein Warteschlangenmodell analysiert. Jedoch erfolgt keine Betrachtung hinsichtlich einer Platzierung von den Servern.

Die Veröffentlichung [127] befasst sich mit der Platzierungsstrategie von Webinhalten auf vordefinierte Server. Insbesondere ist es das Ziel durch gezieltes Management Umleitungen von Anfragen zu vermeiden. Dabei wird auch die Last auf den einzelnen Server mit einbezogen, um die Zugriffszeit und den Lastausgleich zu verbessern. Dennoch ist der Ansatz auf das Netz begrenzt ohne Berücksichtigung von Nutzerverhalten.

Die Konzepte [128, 129] ermöglichen die Empfehlungen des ISP in das System des CDN für eine Kooperation zu integrieren. Dies betrifft vorrangig die Zuweisung von Nutzeranfragen zu Servern und die Bestimmung von Serverstandorten. Der Vorteil besteht darin, dass keine Änderungen am Routing vorzunehmen sind. Der Fokus liegt auf der

Kommunikation und dem Managementprozess zwischen dem CDN Anbieter und dem ISP, wobei dies durch Protokolle wie z. B. ALTO [130] unterstützt wird. Die Arbeit [131] bestätigt die Vorteile durch die Latenz-basierte Serverauswahl und Anfrageumleitung zur Verbesserung der CDN-Performance. Weiterhin bietet der Ansatz [132] eine Strategie zur Speicherung von Inhalten durch den ISP selbst. Dennoch werden keine Informationen zur Lösung der zu Grunde liegenden Probleme im Zusammenhang mit den Zielen der vorliegenden Arbeit beschrieben. Ebenfalls sind dynamische Aspekte hinsichtlich einer analytischen Simulation außer Acht gelassen.

4.4 Ganzheitliche Ansätze

Die bestehenden wissenschaftlichen Arbeiten, welche im direkten Zusammenhang mit der vorliegenden Arbeit stehen, sind in diesem Abschnitt erläutert. Hierbei lässt sich auch der Fortschritt hin zu einem generischen Ansatz erkennen. Der gegenwärtige Entwicklungsstand wird jeweils von den Zielen der vorliegenden Arbeit abgegrenzt.

Für das FLP werden in der Arbeit von Cabrera et al. [37] verschiedene Anwendungsgebiete beschrieben. Durch den gleichzeitigen Bezug auf Logistik und verteilte Rechnersysteme wird eine gemeinsame Problemstellung dieser Bereiche aufgezeigt. Der Gedanke wird in der vorliegenden Dissertation weiter verfolgt und zu einem kombinierten Modell zur Lösung und Optimierung von sich ähnlichen Systemen ausgebaut. Im Rahmen der Arbeit von Cabrera et al. [37] wird für ein Beispiel mit 53 Knoten die Platzierung von Daten in einer Community Cloud als FLP modelliert. Deren Optimierungsansatz ist ausschließlich auf die Minimierung der Hop Anzahl zwischen den platzierten Daten und den Nutzern ausgerichtet. Das Modell besitzt keine gewichteten Knoten und Kanten, sodass keine feingranulare Unterscheidung zwischen verschiedenen Netzverbindungen und Knoten möglich ist. Zur Berechnung von optimierten Platzierungen ist ein heuristisches Verfahren aus Basis von *Iterated Local Search* beschrieben, welches vergleichbare Ergebnisse in einer kürzeren Zeit als der Algorithmus GRASP erreicht. Bei den verwendeten Benchmarks steht die Minimierung der Einrichtungskosten und der operationellen Kosten im Vordergrund. Der Fokus der Arbeit liegt auf der Verbesserung des heuristischen Optimierungsverfahrens und lässt sich somit nicht auf die Problemstellung der vorliegenden Arbeit anwenden.

Mit dem Framework [133] wird für verschiedene Typen von Infrastrukturen ein allgemeiner Ansatz zur Modellierung gegeben. Der Ansatz unterscheidet die zwei Verbindungsarten Nutzer-zu-Nutzer und Nutzer-zu-Zentrum bei der Standortplanung. Dabei werden unter anderem die Topologien Stern, Kreis und Hub sowie Baumstrukturen betrachtet. Insbesondere werden Problemstellungen gemäß dem minimalen Spannbaum analysiert. Die Beschreibung des Modells erfolgt dabei über Formeln hinsichtlich der linearen Programmierung. Trotz des logistischen Bezugs wird die Anwendbarkeit auch für weitere Anwendungsgebiete aufgezeigt und alternative Formulierungen gegeben. Die Bestimmung von Standorten geschieht dabei ohne detaillierte Differenzierung zwischen den Knoten und Verbindungen. Der Ansatz fokussiert auf die Abbildung der grundlegenden Infrastrukturen und bezieht sich nicht direkt auf Transportketten.

In der Analyse [134] wurde die Popularität von geteilten Inhalten bei sozialen Netzen analysiert, insbesondere von Videos auf Twitter. Dabei stellte sich ein geographischer und zeitlicher Zusammenhang heraus. Die Arbeit zeigt durch die Ausbreitung der Information in der Nutzergemeinschaft das Problem der Speicherung von Inhalten auf einem Server des CDN und die notwendige Korrelation von ähnlichen Nutzerinteressen. Insbesondere ist dies durch die Publikation [135] gestützt, wobei gegenwärtige Strategien von CDN Schwächen bei ähnlichen Problemen zur Bereitstellung von nutzergenerierten Inhalten mit hoher QoS und QoE zeigen. Es wird ein Algorithmus zum Management der Anfragen durch eine zentrale Leitstelle auf die einzelnen Knoten des Community-Driven CDN vorgestellt. Das Ziel ist die Minimierung der Betriebskosten bei Erhalt der Servicequalität durch eine angepasste Gruppierung der Nutzer. In Bezug zur vorliegenden Arbeit wird die Problemstellung anhand eines Beispiels vertiefend motiviert. Allerdings ist der Lösungsansatz nur für eine spezifische Anwendung im Bereich CDN und dem Aspekt des Managements von Anfragen nutzbar. Weiterhin wurde in der Arbeit [136] eine Verringerung der Trefferquote der temporären Zwischenspeicherung beobachtet, wenn die klassischen Verfahren für Inhalte aus sozialen Netzen angewendet werden.

Die Arbeiten [36, 137] beziehen sich auf den Anwendungsbereich Rechnernetze und zeigen den Zusammenhang zwischen den konkurrierenden Interessen der Betreiber. In diesem Fall ist dies der Anbieter der Infrastruktur und dem Bereitsteller zur Datenverteilung. Die Analysen demonstrieren, dass die separate Betrachtung der Standortwahl der Server und der Zuweisung der Nutzer mit der Datenübertragung zu suboptimalen Lösungen führt. Ein sequentieller Ansatz ist in der Veröffentlichung [138] gegeben. Somit dienen diese als Ausgangspunkt für ein besseres Verständnis dieser Optimierungsaspekte für die vorliegende Arbeit. Das statische System verwendet dazu keine Nutzerprofile und konzentriert sich ausschließlich auf die Infrastruktur.

Das Dokument [38] beschreibt die Herausforderungen im Bereich des hochperformanten Switchings und Routings. Dabei wird auf die Gemeinsamkeiten von FLP, Hub Location Problem und Rechnernetzen eingegangen. Insbesondere erfolgt eine Motivation die Optimierungsprobleme von Rechnernetzen mit konventionellen Transportproblemen in einem größeren Modell zu kombinieren. Der Fokus dieser Übersicht liegt auf der statischen Optimierung von Netzen mittels Mixed-Integer Linear Programmen hinsichtlich höherer Redundanz und Robustheit, jedoch werden keine direkten Lösungsvorschläge geben oder konkrete Ergebnisse präsentiert. Derselbe Autor entwickelte zudem in einer früheren Arbeit [139] ein lineares Programm für kapazitätsbeschränkte FLP. Jedoch wird keine LP-Relaxation für das Modell angegeben und es weist eine hohe Komplexität auf, wobei dieses für eine effiziente Lösungsberechnung nicht anwendbar ist.

Das in dem Artikel [140] beschriebene Modell bezieht sich zwar auf CCN, umfasst jedoch mehrere Aspekte. Das System ist im Vordergrund auf den Dateninhalten aufgebaut und modelliert diese entsprechend. Mehrere Arten von Daten werden über eine fest vorgegebene Infrastruktur transferiert. Dabei weisen alle eine identische Größe auf, wie auch in anderen Arbeiten [141]. Das Modell umfasst zwei Nutzer, wobei diese einen Datensatz mit der jeweiligen spezifizierten Wahrscheinlichkeit verwenden. Ziel des Ansatzes ist eine Reduktion von Transitgebühren von ISPs [142] in hierarchisch strukturierten Netzen. Auf Grund der spezifischen Ausrichtung des Modells und der

eingeschränkten Ausgangssituation ist dies nicht auf die in dieser Arbeit beschriebene Problemstellung übertragbar.

4.5 Modellierung von Nutzerprofilen und Analyse des Speicherverhaltens in Zentren

In Anbetracht der Problemstellung Nutzerprofile mit in die Berechnungen einzubeziehen, sind dafür im Folgenden wissenschaftliche Arbeiten aufgeführt. Insbesondere sind hierbei geeignete Ansätze zur Adaption und Weiterentwicklung zu identifizieren sowie realistische Parameter zu erhalten.

Im Bereich der Logistik stellt das Buch [143] eine detaillierte Klassifizierung von Standortplanungsproblemen auf. Im Speziellen werden erweiterte Problemstellungen mit variablen Produktionsverhältnissen oder elastischer Nachfrage betrachtet. Dazu sind entsprechende Standort- und Mengenmodelle bei räumlich diversifizierten Marktsituationen aufgestellt. Dabei wird hervorgehoben, dass die Bedürfnisse der Verbraucher zu berücksichtigen sind. Dennoch werden dazu keine konkreten Lösungsvorschläge gegeben oder Ergebnisse präsentiert.

In dem umfassenden Werk [144] für *Supply Chain Management* wird die Notwendigkeit der Kundenorientierung in Zusammenhang mit dem Verbrauchsmanagement bei Transportketten hervorgehoben. Darüber hinaus sind die umfassenden Managementansätze mit vielen Fallstudien bekannter Unternehmen erläutert. Die Kundenausrichtung ist ebenso in [13] beschrieben, wodurch eine angepasste Lösungsermittlung für den Bereich der Logistik motiviert wird. In Bezug auf die explizite Standortplanung von Lagern und die Modellierung von Nutzerprofilen sind allerdings keine Ansätze beschrieben.

Die Arbeit [145] ist spezialisiert auf CCN und *Named Data Networking*. Es stellt ein analytisches Modell für die Speicherung von Inhalten im Cache von Routern mittels Markov-Ketten dar. Hierbei werden Nutzerinteressen in ausstehenden Anfragetabellen aggregiert. Das Modell ist mittels Simulation unter Einsatz der Zipf-Verteilung zur Inhaltspopularität validiert. Die betrachteten Datenpakete sind in einheitlich große Datenpakete eingeteilt. Es erfolgt keine explizite Platzierung von Zwischenspeichern, da bei diesem Ansatz pauschal jeder Router einen Speicher besitzt. Das Modell ist nur auf statische, nicht verändernde Interessen ausgelegt und lässt sich nicht dahingehend dynamisieren. Weiterhin erfolgt in dem Kontext nur die Betrachtung von kleinen Datenpaketen mit kurzfristiger Zwischenspeicherung. Der Fokus der Arbeit liegt in der Analyse der Netzbelastung. Die Tabelle 4.1 zeigt die verwendeten Parameter.

Tabelle 4.1: Parameter der Simulation und der Zipf-Verteilung von [145]

Zipf Parameter α	0.75
Dienste	4200
Speichergröße	1 %; 5 % aller Dienste
Anfragen pro Nutzer	10.000

In einer ähnlichen Arbeit [146] wird ein analytisches Markov-Modell verwendet, um den Einfluss von probabilistischen Caching Verhalten von Routern in CCN auf die Cache Ersetzungsstrategien darzustellen. Hierbei wird die Annahme einer *Zero Download Delay* getroffen. Bei einem Cache miss zu einer Anfrage wird sofort das benötigte Datenfragment in den Cache geladen. Diese realistische Annahme beruht darauf, dass eine Anfrage zufrieden gestellt werden muss, bevor sich die nächste bearbeiten lässt. Da diese Arbeit mit Markov-Ketten arbeitet, lassen sich die Ergebnisse nicht direkt auf dislozierte Systeme übertragen. Weiterhin werden nur die Ersetzungsstrategien Random, *First In First Out* (FIFO) und LRU betrachtet, womit durch hybride Verfahren und moderne Methoden wesentliche Ansätze fehlen. Das Profil der Anfragen an den diskreten Cache wird mit der Zipf-Verteilung generiert wie auch in [39, 146, 147].

In der Arbeit [148] zum Thema *Information-Centric Networking* (ICN) wird eine verbundene Optimierung von Inhaltsreplikationen und domänenübergreifenden Datenverkehr vorgenommen. Ziel der Arbeit ist es den Profit im Bereich vom interdomänen Datenverkehr eines ISP zu erhöhen. Dabei wird der Einfluss auf die Leistungsfähigkeit zur Anfragenbearbeitung vernachlässigt. Die Kosten für den Datenverkehr werden anhand der Popularität des vordefinierten Inhalts berechnet und in einem statischen Ansatz optimiert. Weiterhin ist es nicht von Belang, ob der Inhalt für den eigentlichen Nutzer schnell verfügbar ist. Die Popularität eines Inhalts wird ebenfalls synthetisch über die Zipf-Verteilung erzeugt. Als verifizierte Quelle für diese realistische Annahme wird auf die Anfrageverteilung von Webinhalten verwiesen [149], wobei mehrere Beispiele untersucht wurden [150–154]. Der Parameter α liegt dabei zwischen 0,6 und 1,4 mit der Tendenz zu 0,8. Weiterhin wird davon ausgegangen, dass die Anfragehäufigkeit der verschiedenen angefragten Inhalte eines Nutzers vorab bekannt sind bzw. sich mit unerheblicher Abweichung bestimmen lassen.

Das die Zipf-Verteilung ein valides Modell für Nutzerprofile ist, wurde auch in den Arbeiten [149, 155] nachgewiesen. Hierbei erfolgte eine Analyse von Anfrageverläufen zu Webseiten. Diese Daten entstanden durch sechs akademische Institutionen, indem die Anfragen über einen längeren Zeitraum durch einen Proxy aufgezeichnet wurden. Zudem sind Beispiele aufgeführt, welche die Menge der Verbindungen im Backbone-Netz des Internets analysieren. Daher wird dies auch in der vorliegenden Arbeit zur Modellierung des Nutzerverhaltens angewendet.

Die Arbeit [156] betrachtet persistentes Zwischenspeichern von Inhalten in ICN, wodurch eine Ähnlichkeit zu CDN gegeben ist. Anhand von YouTube und allgemeinen Webserver Datenverkehr wird gezeigt, dass die Nutzung von Repositories die Verfügbarkeit von Inhalten und die Trefferquote im Cache erhöht. Die Häufigkeit von Nutzeranfragen wird entsprechend der betrachteten Szenarien über die parametrisierte Zipf-Verteilung mit $\alpha = 1$ für Web-Inhalte und $\alpha = 2$ für Videos gesteuert, wie auch in [157]. Die Arbeit basiert auf Baumstrukturen, sodass dieser Ansatz sich nicht auf allgemeine Graphenstrukturen übertragen lässt. Zudem wird ausschließlich das Caching mittels modifiziertem LRU betrachtet. Die Arbeit analysiert näher den Einfluss der Größe einzelner Daten. Für Video-Dienste, wie z. B. YouTube gibt es gemäß [158, 159] keine starke positive oder negative Korrelation zwischen globaler und lokaler Popularität. Demgegenüber zeigen mehrere Studien von sozialen Netzdiensten, dass Nutzerinteressen einen hohen örtlichen Bezug

aufweisen [160–164]. Die sozialen Verbindungen und Interessen eines Nutzers lassen sich somit überwiegend auf die Gemeinsamkeiten an dem Umgebungsort gruppieren. Demzufolge weisen Menschen auf Grund deren geografischen Nähe ähnliche Tendenzen beim Zugriff auf bestimmte Dienste auf. In der vorliegenden Arbeit wird diese Eigenschaft berücksichtigt, indem vom ungünstigsten Fall bei der Optimierung auszugehen ist, wobei keine lokale Popularität existiert.

In der Arbeit [165] wird das Speicherverhalten in CCN für Multimedia Daten untersucht. Dafür erfolgt eine Analyse der Trefferrate des Caches in einer zweistufigen Speicherhierarchie, um die notwendige Bandbreite der Netzverbindungen zu reduzieren. Dabei zeigen sich Vorteile von Speichern bei Routern im Bereich von Video-on-Demand. Die Daten werden über eine Zipf-Verteilung mit den Parameter α von 0,8 und 1,2 generiert. Der Ansatz [166] schlägt zur weiteren Optimierung eine Einteilung des Speichers in drei Bereiche von persistente bis flüchtige Inhalte vor. Die verwendeten Modelle basieren auf einer synthetischen Infrastruktur. Diese ist speziell für die Analyse erstellt, sodass dieses sich nicht auf die Problemstellung der vorliegenden Arbeit übertragen lässt.

Die Arbeit [167] präsentiert ein Modell zur proaktiven Speicherung von Inhalten auf den Servern in einem ISP-Netz. Dabei wird auf theoretischer Ebene ein besseres Speicherverhalten gegenüber der Strategie LRU erreicht. Der zentrale Managementansatz bedarf jedoch einen hohen Verwaltungsaufwand. Zudem ist Wissen über die Zugriffspattern notwendig, welche oft auf Grund fehlender Historie nicht vorhanden sind. Diese Techniken lassen sich zusätzlich zu denen in der vorliegenden Arbeit einsetzen, um die Leistungsfähigkeit weiter zu erhöhen.

Zur Vorhersage von Website-Aufrufen von Nutzern bietet das Verfahren über eine Clusteranalyse einen Ansatz [168]. Dazu werden die Anfragen der Nutzer in einer Historie aufgezeichnet. Aus diesen vorausgegangenen Anfragen lässt sich ein Profil erstellen, welches als Grundlage für die Analyse dient. Durch die Cluster von Nutzern mit ähnlichen Profilen lassen sich zukünftige Anfragen vorhersagen. Dieser Ansatz verfolgt dies lediglich für einen Web-Server. Die Idee wird in der vorliegenden Arbeit aufgegriffen und auf ein größeres System skaliert, um die Leistung hinsichtlich der Kooperation von Infrastrukturanbieter und Transportsystem zu verbessern.

4.6 Zusammengefasste Gegenüberstellung bisheriger Arbeiten

Die Tabelle 4.2 stellt die bisherigen Arbeiten in Bezug zu den identifizierten Anforderungen 3.4 gegenüber und zeigt den Stand der Wissenschaft. In der Tabelle ist jeweils nur der erste Autor als Bezeichner der wissenschaftlichen Arbeit genannt. Die Einteilung der Arbeiten erfolgt entsprechend der Abschnitte in diesem Kapitel in die folgenden Bereiche: Ganzheitlich, Logistik, HaS, Rechnernetze, Nutzerprofil und Algorithmen. Diese sind jeweils chronologisch sortiert.

Die Abkürzungen in der Tabelle haben folgende Bedeutung:

- FLP ... Facility Location Problem
- RN ... Rechnernetze
- SC ... Supply Chain
- HaS ... Hub and Spoke
- ICN ... Information Centric Networking
- CCN ... Content Centric Networking
- CDN ... Content Delivery Network
- ✓ ... Anforderung des Kriteriums wird erfüllt
- o ... Anforderung des Kriteriums wird nur teilweise erfüllt oder trifft nur indirekt zu
- - ... Anforderung des Kriteriums wird nicht berücksichtigt bzw. hat keine Auswirkung

In der Übersicht ist ersichtlich, dass bisher kein Ansatz in dessen Eigenschaften den Anforderungen in deren Gesamtheit gerecht wird. Insbesondere fehlt es an Lösungsvorschlägen für einen generischen Modellierungsansatz mit multikriterieller Optimierung zur Einbeziehung des Nutzerverhaltens. Dies ist mutmaßlich auf die Komplexität des Gesamtzusammenhangs und das Fehlen von Lösungsmöglichkeiten für Teilprobleme zurück zu führen. Zudem gab es bisher nur wenige umfassende Aufzeichnungen und Analysen von Nutzerverhalten gemäß Profiling. Erst mit den vorausgegangen Arbeiten der letzten Jahre wurden entsprechende Grundlagen hinsichtlich der Modellierung von Transportketten und Nutzerprofilen geschaffen. Weiterhin ermöglicht die gestiegene Leistungsfähigkeit der Technik die Verarbeitung von größeren Datenmengen und Berechnungen aufwendiger Optimierungsalgorithmen.

Die Abbildungen 4.1 und 4.2 verdeutlichen die gegenwärtigen Defizite in den fokussierten Anwendungsgebieten. Dies wird insbesondere durch die markierten Eigenschaften hervorgehoben, wobei deren Bedeutung mit den entsprechenden Bewertungskriterien der Anforderungsanalyse übereinstimmt. Bei den Wertschöpfungsketten in der Logistik zeigt sich Bedarf bei der Leistung als auch bei der Kundenorientierung. Im Bereich von CDN besteht Verbesserungspotential in der Anwendbarkeit von Modellen [41].

Tabelle 4.2: Übersicht bisheriger Arbeiten bewertet entsprechend der Vergleichseigenschaften (✓ bedeutet erfüllt; o zeigt Schwächen in der Anwendbarkeit an; - steht für nicht zutreffend)

Literatur	Anwendungsgebiet	Anzahl der Zentren	Optimierte Platzierung der Zentren	Wechselseitige Abhängigkeit	Differenzierung von Verbindungspfaden und Nutzern	Last auf die Infrastruktur	Zuweisung der Nutzer	Nutzerprofile	Bereitstellung des Dienstes	Speicherkapazität der Zentren	Simulation
Jiang 2009 [137]	CDN	-	✓	-	o	✓	-	-	-	-	-
Contreras 2012 [133]	FLP	o	✓	✓	-	-	o	-	-	-	-
Cabrera 2014 [37]	FLP/RN	-	o	o	-	-	o	-	-	-	-
Hu 2014 [134]	CDN	-	-	-	-	-	o	o	✓	o	-
Kamiyama 2014 [140]	CCN	-	-	-	o	-	o	o	o	-	-
Rodrigues 2014 [36]	CDN	o	o	-	-	-	o	-	o	-	-
Papadimitriou 2016 [38]	FLP/RN	-	o	o	-	o	-	-	o	-	-
Francis 1983 [50]	FLP	o	✓	-	-	o	o	-	-	-	-
Khuller 2000 [71]	FLP	o	✓	-	-	o	-	-	-	-	-
Jang 2002 [83]	SC	o	-	-	-	-	o	-	✓	o	-
Alicke 2005 [84]	SC	o	o	-	o	-	-	-	✓	o	-
Snyder 2006 [77]	FLP	o	✓	-	-	-	✓	o	-	-	-
Vahrenkamp 2007 [85]	FLP	-	✓	-	✓	-	o	-	-	-	-
Zeng 2007 [81]	FLP	-	✓	-	-	-	o	o	-	-	-
Qin 2008 [82]	FLP/SC	-	✓	-	o	-	o	-	o	-	-
Xi-feng 2008 [75]	FLP	o	✓	-	o	-	✓	-	-	-	-
Zhui 2010 [73]	FLP/SC	-	-	-	o	o	-	-	✓	-	✓
Williamson 2011 [79]	FLP	-	✓	-	o	-	-	-	-	-	-
Daskin 2013 [74]	FLP	o	✓	o	-	✓	o	-	-	-	-
Mayer 2001 [88]	HaS	o	o	✓	-	✓	o	-	-	o	-
Owsinski 2015 [89]	HaS	-	o	✓	✓	o	o	-	-	-	-
Ribeiro da Cunha 1997 [98]	CDN	-	-	-	-	o	-	o	✓	-	o
Krishnan 2000 [56]	CDN	o	✓	-	-	o	-	-	o	o	o
Jamin 2001 [92]	CDN	o	o	-	-	o	o	-	-	-	o
Qiu 2001 [127]	CDN	-	o	-	-	o	✓	-	✓	-	o
Buyya 2008 [114]	CDN	-	✓	-	-	o	✓	-	o	✓	-
Andreica 2009 [99]	CDN	-	✓	-	-	✓	-	-	-	-	-
Wang 2010 [106]	CDN	-	-	-	-	o	-	o	o	✓	o
Carofiglio 2011 [90]	CCN	-	-	-	-	-	-	✓	o	✓	o
Dahlmann 2011 [119]	CDN	-	-	-	-	o	-	-	✓	o	-
Li 2011 [113]	CDN	-	o	-	-	o	✓	-	o	✓	-
Frank 2013 [128]	CDN	o	o	-	-	-	-	o	o	-	-
Inaba 2015 [94]	ICN	-	-	-	-	-	o	o	✓	-	o
Mobasher 2000 [168]	CDN	-	-	-	-	-	o	✓	-	-	-
Abu 2015 [145]	CCN	-	-	-	o	✓	-	✓	-	-	o
Feng 2015 [148]	ICN	-	-	-	-	o	-	✓	o	-	-
Tarnoi 2015 [146]	CCN	-	-	-	-	-	-	✓	o	o	o
Kleinberg 2005 [65]	*k*-center	o	✓	-	-	-	-	-	-	-	-

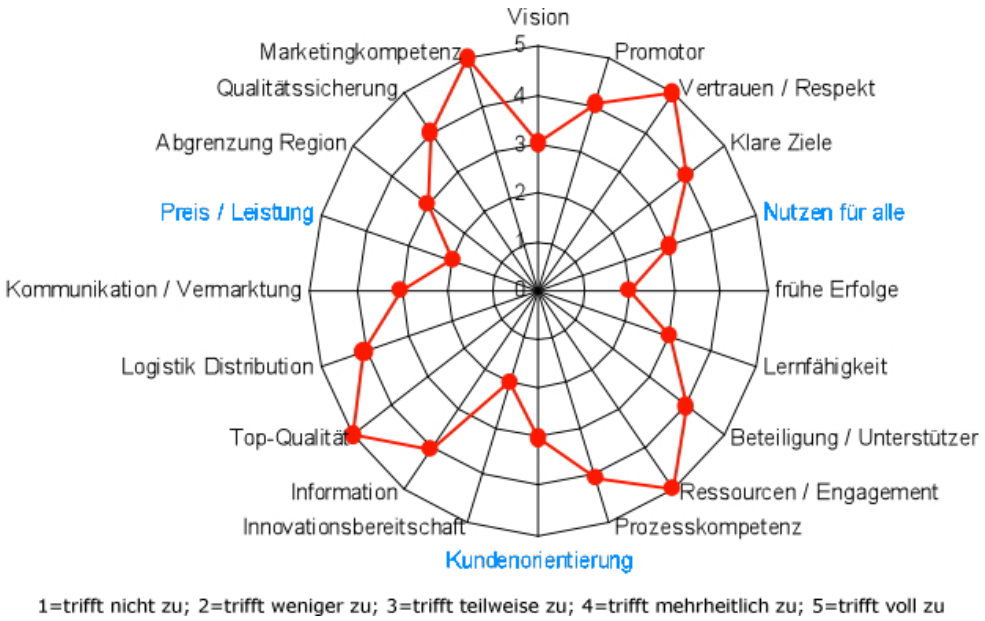


Abbildung 4.1: Bewertung des Stands der Technik anhand von Erfolgsfaktoren im Bereich von FLP (Vgl. [169, 170])

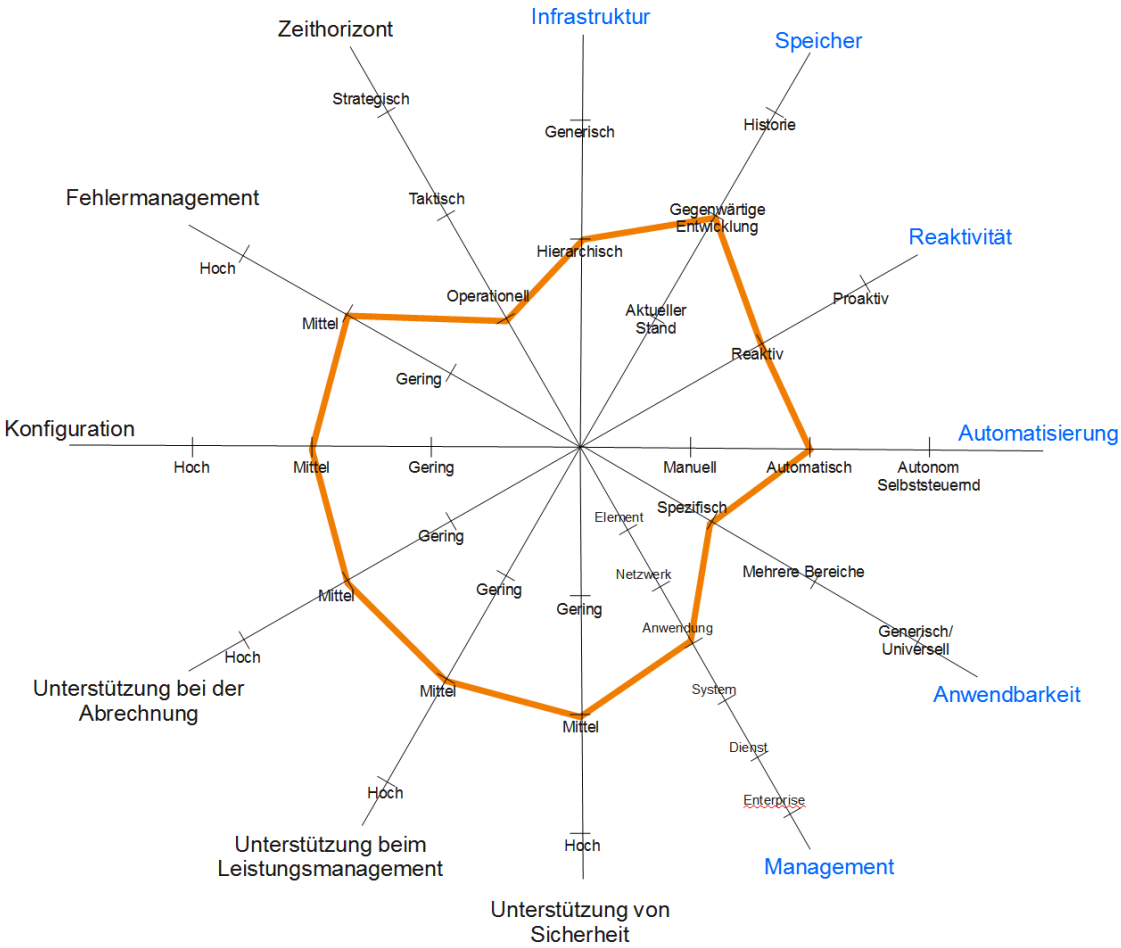


Abbildung 4.2: Bewertung von Modellen anhand von Kriterien zur Kooperation im Bereich von CDN (Vgl. [171])

4.7 Limitierungen derzeitiger Ansätze

Wie aus dem Stand der Wissenschaft in Kapitel 4 ersichtlich, gibt es mehrere, vergleichbare Problemstellungen mit ähnlichen Anforderungen innerhalb einer Domäne. Der Fokus aktueller Ansätze beschränkt sich auf eine Fall-spezifische und Domänen-spezifische Modellierung im jeweiligen Anwendungsgebiet. Eine Ausnahme bildet die Arbeit von Cabrera et al. [37], welche die Problemstellung der Standortwahl aus beiden Anwendungsgebieten abstrahiert betrachtet. Unter anderem wird daher ein generischer Ansatz als zielführend erachtet [38]. Die Optimierung des Problems erfolgt bei derzeitigen Ansätzen hinsichtlich nicht allen identifizierten Zielkriterien für eine übergreifende Lösung. Vor allem wird bei älteren Modellen eine Analyse nur bezüglich eines Zielkriteriums vorgenommen. Insbesondere ergibt sich dadurch ein begrenzter Lösungsansatz für konkurrierende Interessen von kooperativen Dienstanbietern. Somit beschränken sich die Ansätze auf eine lokale Sichtweise der Optimierung. Bisherige Verfahren setzen vorrangig auf eine statische Analyse des Systems und vernachlässigen die dynamischen Aspekte. Das Mittel der Simulation zur Analyse von operativen Effekten wird nur unzureichend eingesetzt, insbesondere im Bereich von CDN. Darüber hinaus gibt es nur eine eingeschränkte Modellierung des Verhaltens von Nutzern gemäß einem Interessen-Profil vor allem im Bereich FLP und verteilter Systeme. Dies führt zu einer Simulation des Systems, welches nicht das individuelle Nutzerverhalten berücksichtigt. Somit ergibt sich bei aktuellen Modellierungs- und Optimierungsansätzen eine begrenzte Beachtung von zusammenhängenden Auswirkungen während der Laufzeit. Dies limitiert die Lösungsberechnung hinsichtlich eines kooperativen und ganzheitlichen Ergebnisses.

4.8 Referenzalgorithmen zur Lösungsfindung in Teilbereichen

Zur Lösung von Teilproblemen der Aufgabenstellung gibt es in der Literatur mehrere etablierte Verfahren. Diese werden als Referenz zur Lösungsfindung mit herangezogen und zum Leistungsvergleich genutzt. Dies gilt insbesondere für die in dieser Arbeit entwickelten Platzierungsverfahren für die Zentren gemäß dem *k*-center Problem. Darüber hinaus erfolgt eine spezifische Adaption der Verfahren an die Problemstellung.

Zu einer Optimierungsaufgabe p sei L(p) der zugehörige Lösungsraum. Die Güte zu jeder möglichen Lösung $r \in L(p)$ ist als d(r) definiert. Weiterhin wird die Güte der optimalen Lösung als d* bezeichnet. Die Optimierungsalgorithmen suchen nach einer Lösung $r \in L(p)$, welche d* ist oder dieser möglichst nahe kommt. Allerdings ist für heuristische Optimierungsverfahren bewiesen, dass diese lediglich eine beweisbare Lösungsgüte von $d(r) = 2 \cdot d^*$ garantieren können. Der Beweis ist im Anhang C beschrieben.

Folgende Optimierungsansätze werden als Referenz verwendet:

- Monte Carlo
- Zwei-Approx (auch *Farthest-First* genannt)
- Greedy

- GRASP
- Partitionierende Clusterverfahren (Lloyd, MacQueen, K-Means++)
- Integer Linear Programming (CPlex, Gurobi)
- Frameworks (ELKI, WEKA, Serein, MOEA)

4.8.1 Monte Carlo

Die Monte Carlo Methode ist ein Verfahren aus der Stochastik und verfolgt keine spezifische Strategie zu Lösung eines Problems [172]. Es basiert auf dem Zufall und macht sich das Gesetz der großen Zahlen zu Nutze. Die Monte Carlo Methode wählt in Abhängigkeit der Platzierungsbedingung zufällige Punkte innerhalb des Lösungsraumes aus. Bei der knotenbasierten Platzierung werden aus den vorgegebenen Positionen beliebige entsprechend der definierten Anzahl von Zentren ausgesucht. Bei einer freien Platzierung der Zentren werden die Standorte innerhalb der durch die Nutzer aufgespannten Bounding Box willkürlich ausgewählt. Eine mehrfache Selektion einer Position ist dabei auszuschließen. Entsprechend der Wiederholungsanzahl erfolgt eine mehrfache Generierung von Lösungen, wobei die Beste ausgegeben wird.

4.8.2 Zwei-Approx

Der Zwei-Approx Algorithmus gehört zu den Cluster-Algorithmen und bestimmt für eine vorgegebene Menge an Knotenpunkten die definierte Anzahl k an zentralen Standorten [173]. Dabei erfolgt eine sequentielle Platzierung der Zentren an den Standorten von Nutzern, welche jeweils am weitesten von den Zentren entfernt sind. Durch die so weit wie nur mögliche, von einander entfernte Platzierung wird der Algorithmus auch als *Farthest-first* bezeichnet [174, 175]. Dieser ist durch die stringente Platzierung von Zentren an vorgegebenen Standorten nur für die Nebenbedingung der knotenbasierten Platzierung anwendbar. Die Funktionsweise ist wie folgt spezifiziert:

1. Wähle zufällig einen Knoten aus, dessen Position das erste Zentrum ist.
2. Berechne für alle Knoten den Zielfunktionswert, die Entfernung zum am Besten gelegenen Zentrum.
3. Wähle den Knoten aus, der den schlechtesten Zielfunktionswert aufweist als die Position für das nächste Zentrum.
4. Wiederhole die zwei vorherigen Schritte bis die spezifizierte Anzahl an Zentren erreicht ist.

Der Algorithmus ist zwei-approximierend und hat eine polynomielle Laufzeit [78, 176]. Ein Approximationsalgorithmus löst Optimierungsaufgaben näherungsweise. Im Allgemeinen garantiert ein β-approximierender Algorithmus mit der Approximationsgüte β, dass die gefundene Lösung $r < \beta \cdot d^*$ für eine Optimierungsaufgabe ist. Dazu muss keine Kenntnis oder Abschätzung über d^* vorliegen, sodass der Algorithmus die Leistung ohne Wissen

über das globale Optimum garantiert. In Bezug auf die Problemstellung gewährleistet der Zwei-Approx Algorithmus Lösungen d gemäß Gleichung 4.1.

$$d \leq 2 \cdot Optimum \tag{4.1}$$

Die maximale Entfernung eines Nutzers zu seinem am Besten gelegenen Zentrum ist somit kleiner gleich als das Doppelte vom Maximum wie bei der optimalen Lösung [54]. Durch die bewiesene Approximationsgüte lässt sich das globale Optimum d^* abschätzen, sodass die Ergebnisse des Algorithmus als Referenz dienen.
Die Lösungen variieren durch die zufällige Initialisierung des ersten Zentrums und der zufallsbedingten Auswahl bei gleichen Zielfunktionswerten. Dennoch gilt stets die zwei-approximierende Güte, sodass sich durch die Umstellung der Gleichung 4.1 ein theoretisches Limit für das Optimum entsprechend der Gleichung 4.2 ermitteln lässt. Zur Bestimmung des theoretischen Optimums werden dafür anstatt des Durchschnittes die pessimalen Werte von den wiederholten Ausführungen verwendet. Die Gleichung gilt hinsichtlich der maximalen Entfernung von den Knoten zum nächstgelegenen Zentrum.

$$d \leq 2 \cdot Optimum_{real} \implies Optimum_{theoretical} \geq \frac{1}{2} \cdot d_{max} \tag{4.2}$$

Die Abbildung 4.3 zeigt die ersten beiden Iterationen vom Ablauf des Algorithmus. Auf der linken Seite wird das erste Zentrum zufällig platziert. In der nächsten Iteration auf der rechten Seite wird der am weitesten entfernte Knoten selektiert.

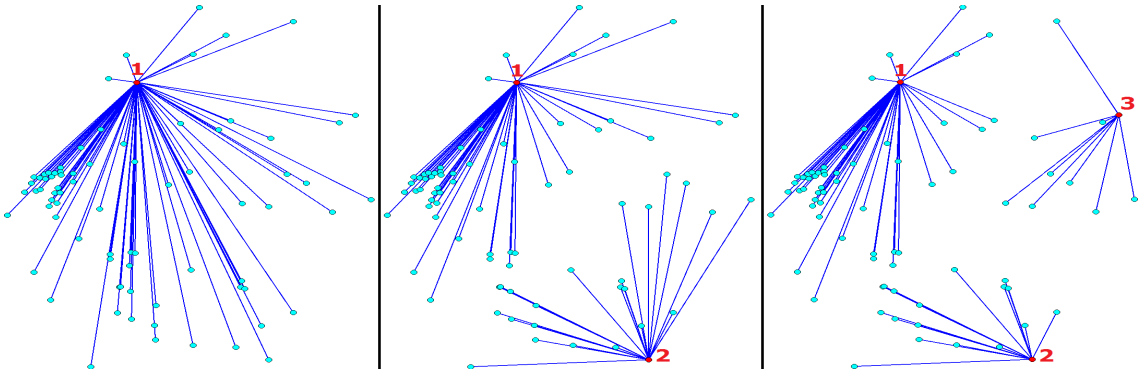


Abbildung 4.3: Die ersten drei Iterationen des Zwei-Approx anhand des Beispielszenarios Deutschland aus der Logistik

4.8.3 Greedy

Die Greedy Strategie nimmt wie auch der Zwei-Approx Algorithmus eine iterative Platzierung der zentralen Knotenpunkte vor [92, 112]. Gemäß der Literatur führt die Strategie eine Platzierung von Zentren nur an vorgegebenen Positionen durch.

In jeder Iteration platziert der Algorithmus ein Zentrum. Ein einmal platziertes Zentrum bleibt an dessen Standort bestehen und wird nicht mehr verschoben. In einer Iteration testet Greedy immer alle verbleibenden Möglichkeiten zur Platzierung. Bei jeder Evaluierung werden die Nutzer entsprechend deren nächstgelegenen Zentrum

neu zugeordnet. Die Gruppierung der Nutzer berücksichtigt dabei stets alle bereits platzierten Zentren. Anschließend wählt die Strategie die Position aus, welche die größte Verbesserung in Anbetracht des Optimierungskriteriums vom Gesamtszenario erbringt. Dieser Prozess wird solange wiederholt, bis die spezifizierte Anzahl an Zentren erreicht ist.

Der Berechnungsprozess verwendet zur Bestimmung der nächsten Position eines Zentrums eine Entfernungsmatrix . In dieser werden die Zielfunktionswerte von jedem Nutzer zu jedem Zentrum und jeder möglichen Position eines weiteren Zentrums gespeichert. In der einfachen Version des Algorithmus wird die Matrix nicht nach jeder Wahl eines Zentrums aktualisiert [177]. Zur weiteren Verbesserung einer Lösung lässt sich zusätzlich sogenanntes Backtracking einsetzen. Dabei überprüft der Algorithmus, ob ein bereits zuvor platziertes Zentrum sich besser platzieren lässt oder entfernt werden kann. Entsprechend der definierten Anzahl δ von Backtracking Schritten werden vorausgegangene Platzierungsentscheidungen gegebenenfalls revidiert und der Algorithmus erhält die Bezeichnung δ-Greedy.

Tabelle 4.3 zeigt die Entfernungsmatrix für ein Beispiel. Bei symmetrischen, ungerichteten Zielfunktionen muss nur die obere oder untere Diagonalmatrix berechnet werden. In dem Beispiel wird die Position vom Knoten N3 als nächstes Zentrum ausgewählt. Durch die vollständige Suche in jeder Iteration ergibt sich ein höherer Berechnungsaufwand als beim Zwei-Approx Algorithmus, wobei Greedy ebenfalls eine zwei-approximierende Lösungen erreicht [65].

Tabelle 4.3: Beispiel-Matrix des Greedy Algorithmus mit 5 Knoten

	N1	N2	N3	N4	N5	$\mathbf{max}_{\text{Zeile}}$
N1	-	5	10	13	8	13
N2	5	-	4	17	9	17
N3	10	4	-	12	6	**12**
N4	13	17	12	-	15	17
N5	8	9	6	15	-	15

4.8.4 GRASP

Eine Erweiterung des Algorithmus Greedy ist *Greedy Randomized Adaptive Search Procedures* (GRASP) [178]. Dies ist eine Meta-Heuristik, welche mehrmals beginnt und somit verschiedene Anfangssituationen als Initialisierung verwendet, um anschließend eine lokale Optimierung mittels Greedy durchzuführen [179, 180]. Die Platzierung erfolgt wie auch bei Greedy nur an vorgegebenen Positionen. In der anfänglichen Initialisierungsphase wird eine gültige Lösung durch eine zufallsgesteuerte und gewichtete Strategie erzeugt. Dazu evaluiert der Algorithmus in jeder Iteration alle Möglichkeiten und vergibt diesen entsprechend dem zu erwartenden Zielfunktionswert eine Wahrscheinlichkeit gemäß Gleichung 4.3.

$$\rho = \min_{Zielwert} + \gamma \cdot \left(\max_{Zielwert} - \min_{Zielwert} \right) \tag{4.3}$$

Der Wert für γ ist entsprechend der Problemstellung zwischen 0 und 1 zu spezifizieren. Anschließend erfolgt auf Grundlage der beeinflussten Wahrscheinlichkeiten eine zufallsbasierte Auswahl. Dabei wird darauf geachtet, dass an einem Standort keine mehrfache Platzierung erfolgt. Der Vorgang wiederholt sich solange, bis die vorgegebe Anzahl an Zentren erreicht wurde. Anschließend erfolgt in einem iterativen Prozess eine lokale Suche in den benachbarten Lösungen bis zum lokalen Optimum. Dazu wird in einem lokalen Schritt zufällig eins der Zentren ausgewählt, aus der Lösungsmenge entfernt und mittels Greedy eine neue Position zugewiesen. Zur Steigerung der Effizienz wird verhindert, dass ein Zentrum zweimal hintereinander ausgewählt wird. Die beste Gesamtlösung wird dabei als Ergebnis zurückgegeben.

4.8.5 Partitionierende Clusterverfahren

Die im Folgenden beschriebenen Clusterverfahren gehören zu der Kategorie der K-Means Algorithmen. Die grundlegende Idee basiert darauf die Nutzer zu k Clustern zusammen zu fassen. Für jeden Cluster wird ein Zentrum definiert. Da die iterative Platzierung neuer Zentren durch die Neuzuordnung der Nutzer die Standorte bereits platzierter Zentren beeinflussen, erfolgt bei diesem Ansatz stets eine gleichzeitige Bestimmung der Standorte für alle Zentren. Eine Platzierungsänderung eines Zentrums hat somit direkten Einfluss auf die anderen Zentren. Dadurch wird unter anderem Backtracking unnötig und steht im Gegensatz zu den iterativen Ansätzen.

Es existieren mehrere Algorithmen zum Clustern von Knoten mit einer vorgegebenen Anzahl an Zentren nach diesem Prinzip. Die verschiedenen K-Means Algorithmen unterscheiden sich hauptsächlich bei der Initialisierung der Startlösung. Die Erläuterung erfolgt anschließend bei der konkreten Beschreibung des jeweiligen Algorithmus. Nach der Initialisierung wird jedem Nutzer dasjenige Zentrum zugeordnet, welches den besten Zielfunktionswert ergibt. Bei der Platzierung ist dies das Zentrum mit der geringsten Entfernung entsprechend der Gleichung 4.4.

$$G_j^t = v_i : d(v_i, k_u) \leq d(v_i, k_{u*}) \text{ für alle i} \in 1..n \tag{4.4}$$

Für jeden Cluster von Knoten, welcher einem Zentrum zugeordnet ist, wird anschließend eine aktualisierte Position berechnet. Dabei ergibt sich die Position des Zentrums in jeder Iteration t aus dem Mittelwert der Standorte aller zugeordneten Nutzer v_i. Der geometrische Schwerpunkt von allen Knoten einer Gruppe G_j wird entsprechend der Gleichung 4.5 berechnet. In jeder Iteration werden die Knoten erneut dem jeweiligen nächstgelegenem Zentrum zugewiesen. Dieser Prozess wird solange wiederholt, bis sich die Standorte der Zentren nicht mehr ändern. Es ist bewiesen, dass diese Verfahren immer terminieren [181]. Die Leistung ist durch die zufällig gewählten Positionen von Zentren beeinflusst und findet nicht notwendigerweise eine optimale Lösung. Um diesen Effekt zu minimieren, wird der Algorithmus mehrfach mit unterschiedlichen Initialisierungen ausgeführt. Falls es bei der iterativen Optimierung zu dem Fall kommt, dass ein Zentrum keinen Nutzer mehr hat, dann ist dieses grundlegend neu zu platzieren. Dazu wird es zu dem Nutzer mit der größten Entfernung zu einem Zentrum gesetzt.

In Abhängigkeit der Nebenbedingungen sind die verschiedenen Verfahren entsprechend daran anzupassen.

$$k_u^{t+1} = \frac{1}{\left|G_j^t\right|} \cdot \sum_{v_i \in G_j^t} k_u \tag{4.5}$$

Llyod-Algorithmus

Der Algorithmus von Llyod [182] wird als der ursprüngliche K-Means Algorithmus angesehen und ist auch als Voronoi Relaxation [183] bekannt. Vor der Initialisierung ermittelt der Algorithmus die Bounding Box von den gegebenen Nutzern und bestimmt somit den Lösungsraum. Anschließend erfolgt die Platzierung der Zentren innerhalb der Grenzen der Bounding Box. Dazu erzeugt der Algorithmus entsprechend der vorgegebenen Anzahl k Zentren mit zufälligen Werten. Nach der Initialisierung werden die Standorte der Zentren iterative verbessert, wobei die zuvor beschriebene K-Means Strategie zum Einsatz kommt.

MacQueen-Algorithmus

Der MacQueen-Algorithmus [184] ist eine vereinfachte Form des Lloyd Algorithmus. Dieser verwendet bei der Initialisierung eine Menge vorgegebener Standorte, aus denen zufällig die definierte Anzahl k Zentren ausgewählt wird. Daher sind dem Algorithmus spezifische Positionen für die Initialisierung vorzugeben. In Anbetracht des Anwendungsszenarios lassen sich dafür die Standorte der Nutzer verwenden. Anschließend erfolgt die iterative Zuweisung der Nutzer zu den Zentren und die Aktualisierung des Standortes gemäß der K-Means Strategie.

K-Means++ Algorithmus

Eine weitere Variation ist der K-Means++ Algorithmus [185, 186]. Hierbei wird die zufällige Auswahl der Startpunkte bei der Initialisierung über Wahrscheinlichkeiten beeinflusst. Ziel ist dabei bestimmte Positionen zu begünstigen, wodurch in der anschließenden Optimierung im Mittel bessere Ergebnisse entstehen. Dieses Verhalten wurde in den Publikationen mathematisch nachgewiesen. Die Auswahl der Standorte für die Zentren erfolgt ausschließlich bei den vorgegebenen Platzierungsmöglichkeiten. Die Position des ersten Zentrums wird wie bisher komplett zufällig bestimmt. Für alle weiteren Zentren wird die Wahrscheinlichkeit zur Auswahl des Standortes verzerrt. Die Auswahlwahrscheinlichkeit für ein Zentrum entspricht in jeder Iteration direkt proportional dem quadratischen Wert der Zielfunktion gemäß Gleichung 4.6. Der Wert ergibt sich dabei stets in Relation zu bereits platzierten Zentren.

$$\rho(k_u) = d_u^2 \,/\, \sum_{i=1}^{n} d_i^2 \tag{4.6}$$

Der zu erwartende Zielfunktionswert E für ein Ergebnis entspricht der Gleichung 4.7

$$E \leq 8(ln\ k + 2) \cdot Optimum \tag{4.7}$$

Darüber hinaus gibt es insbesondere für den Anwendungsbereich des Data Minings weitere Optimierungsverfahren [187–189], die nach dem K-Means Prinzip funktionieren und sich in der Initialisierung unterscheiden.

4.8.6 Integer Linear Programming

Bei der Linearen Programmierung (LP) wird das Problem mittels mathematischen Gleichungen beschrieben und mit Hilfe von LP-Solvern gelöst [65, 190, 191]. Da bei der Lösungsberechnung ausschließlich ganzzahlige Werte anzunehmen sind, lässt sich das Problem als Integer Linear Programming (ILP) genauer spezifizieren. Als Solver kommen hierbei GUROBI [192] und CPLEX [193] zum Einsatz, welche selbst bei unimodularen Matrizen in polynomieller Zeit Lösungen ermitteln. Beide Ansätze verwenden eine jeweils eigens abgewandelte Form des Simplex Algorithmus von Gorge Danzig [194]. Die native Ergebnisberechnung löst ein ILP Problem durch das Entfernen der Bedingung, dass die Variablen ganzzahlig sein müssen. Das entstehende LP-Modell wird dementsprechend die LP-Relaxation vom ILP genannt. Dies ist ein Standardverfahren zur Näherungsweisen Lösung eines NP-Problems. Durch entsprechendes Runden der Lösungsvariablen von der LP-Relaxation lässt sich gegebenenfalls eine Lösung für das ILP-Modell erreichen. Jedoch muss die entstandene Lösung nicht optimal sein, noch überhaupt zu einer Lösung für ILP führen. Die Integritätslücke gibt das maximale Verhältnis von einer Lösung des ILP und dessen Relaxation an [195].

Das folgende ILP-Modell entspricht der nicht kapazitiven Problemstellung und ist an das jeweilige Zielkriterium anzupassen [62, 85, 196–199]. Dieses wird unter anderem in dem weiteren Anwendungsgebiet der Geolokalisation zur Bestimmung von zentralen Netzknoten für Zeitmessungen eingesetzt [29]. Das Modell garantiert eine Lösung mit dem Zielfunktionswert $d \leq 2 \cdot Optimum$ und besitzt somit eine Integritätslücke von zwei [200, 201]. Das Verfahren lässt sich zur Platzierung von Zentren an vorgegebenen Standorten verwenden und dient somit als Referenz zu den anderen Algorithmen. Für Spezialfälle gibt es erweiterte Modelle mit kleineren Integritätslücken, wobei ein kubischer oder noch höherer Rechenaufwand notwendig ist [202, 203]. In einer Erweiterung lassen sich auch die Entfernungen zu weiteren Zentren berücksichtigen, sodass ein fehlertolerantes System entsteht [204]. Bei Problemstellungen mit vielen Variablen wird von der LP Nutzung auf Grund des hohen Berechnungsaufwandes abgewichen und heuristische Verfahren eingesetzt [119].

Die Beschreibung des Modells ist wie folgt:

- Der Knotenpunkt v_i ist als Zentrum k_u ausgewählt (Wert = 1) oder nicht (Wert = 0), Gleichung 4.8.
- Die Anzahl k der ausgewählten Standorte für Zentren darf kleiner gleich der vorgegeben Zahl sein, Gleichung 4.9.

- Eine Verbindung zwischen zwei Kontenpunkten $d(v_i, v_j)$ ist für die Evaluation zu beachten (Wert = 1) oder nicht (Wert = 0). Dies entspricht der Zuweisung eines Nutzers zu einem definierten Zentrum, Gleichung 4.10.
- Jeder Nutzer muss genau eine Verbindung zu einem Zentrum haben, Gleichung 4.11.
- Ein Nutzer kann nur von einem Knotenpunkt versorgt werden, wenn dieser als Zentrum ausgewählt wurde, Gleichung 4.12.
- Zielfunktion: In Abhängigkeit der Problemstellung ist zwischen den Gleichungen für k-median 4.13 oder k-center 4.14 und 4.15 zu unterscheiden.

$$v_i \in 0, 1, \ \forall i \tag{4.8}$$

$$\sum_{i=0}^{n} v_i \leq K \tag{4.9}$$

$$e(v_i, v_j) \in 0, 1, \ \forall i, j \tag{4.10}$$

$$\sum_{i=0}^{n} e(v_i, v_j) = 1, \ \forall j \tag{4.11}$$

$$e(v_i, v_j) \leq v_j, \ \forall i, j \tag{4.12}$$

Zielfunktion k-median:
Das Modell für k-median verwendet die folgende Zielfunktion [205, 206]. Mit $d(v_i, v_j)$ wird die Entfernung zwischen zwei Knotenpunkten angegeben. Dies ist unabhängig von der gewählten Entfernungsfunktion. Das Ziel ist die Minimierung der Summe der Distanzen.

$$\sum_{i=0}^{n} d(v_i, v_j) \cdot e(v_i, v_j), \ \forall j \ \rightarrow min \tag{4.13}$$

Zielfunktion k-center:
Das Modell für k-center verwendet die folgende Zielfunktion [207]. Mit $d(v_i, v_j)$ wird die Entfernung zwischen zwei Knotenpunkten angegeben. Das Ziel ist die Minimierung der Distanzen.

$$d(v_i, v_j) \cdot e(v_i, v_j) \leq z, \ \forall \ i, j \tag{4.14}$$

$$z \rightarrow min \tag{4.15}$$

4.8.7 Frameworks für Algorithmen

Für das Clustern von Daten gibt es bereits Frameworks, die neben den zuvor genannten Algorithmen noch weitere Verfahren implementiert haben. In der Programmiersprache Java sind für die Problemstellung ELKI und WEKA zutreffende Frameworks zum Data Mining. Unabhängig vom Anwendungsgebiet gibt es weiterhin Meta-Heuristiken zur näherungsweisen Lösung von Optimierungsproblemen. Hierbei bieten sich die Frameworks SEREIN und MOEA an.

Framework ELKI

ELKI (Environment for Developing KDD-Applikations Supported by Index-Structures) ist eine Open-Source Software für Data-Mining [208]. Diese steht unter der *GNU Affero General Public License*. Das Framework wird von der Ludwig-Maximilians-Universität München in Java programmiert und befindet sich noch im Entwicklungsstadium. Der Anwendungsbereich fokussiert auf die Bereitstellung von Algorithmen zur Cluster-Analyse. Dabei werden verschiedene Datenstrukturen sowie eine grafische Oberfläche zur Berechnung und Auswertung angeboten. ELKI bietet eine Vielzahl von parametrisierbaren Optimierungsalgorithmen. Die angebotenen Verfahren sind im Folgenden aufgelistet.

- K-Means MacQueen
- K-Means Lloyd
- K-Means Hybrid Lloyd - MacQueen
- K-Means Batched Lloyd
- Parallel Lloyd K-Means
- K-Medians Lloyd
- K-Medoids PAM (Partitioning Around Medoids) [209]
- K-Medoids EM
- Single Assignment K-Means

Da es auf Grund des verwendeten Entwicklungsstatus von Version 0.6.5 noch zu Veränderungen in der API kommt, wird eine externe Ansteuerung über die Kommandozeile empfohlen [210]. Dazu muss ein Szenario entsprechend dem Format von ELKI exportiert und importiert werden. Dies führt zu erheblichen Verzögerungen bei der Berechnung, sodass die Algorithmen für diese Arbeit adaptiert wurden.

Framework WEKA

WEKA (Waikato Environment for Knowledge Analysis) ist eine Open-Source Software für Machine-Learning und Data-Mining [211]. Diese steht unter der *GNU General Public License*. Das Framework wird von der Waikato Universität in Neuseeland in Java programmiert. Hierbei handelt es sich um eine ausgereifte Software, welches eine direkte Nutzung des Quellcodes ermöglicht. Es werden verschiedene Aufgabentypen hinsichtlich von Data-Mining unterstützt, sodass eine umfassende Analyse von Daten möglich ist. Die bereitgestellten Algorithmen sind weniger umfangreich als bei ELKI.

Framework SEREIN

Zum weiteren Vergleich wird das in Java implementierte Framework SEREIN [212, 213] eingesetzt. Damit lassen sich verschiedene von der Natur inspirierte Verfahren und

evolutionäre Algorithmen abbilden. Die standardmäßig angebotenen Strategien sind: Genetischer Algorithmus [214] und Simulated Annealing [215]. Als Referenz wird vorrangig der Genetische Algorithmus mit den folgenden Parametern eingesetzt:

- Genetische Darstellung: Vektor mit reellen Zahlen r mit 0 <= r <= 1.0 zur Abbildung der Koordinaten von Zentren
- Variationsoperator: Creep Mutation mit Standardabweichung und ohne Rekombination
- Population/Generation: 25 Individuen über 80 Generationen
- Selektion: Tournament mit der Größe von 2

Die Parameter der evolutionären Algorithmen wurden über Meta-Optimierung experimentell bestimmt.

Framework MOEA

Das Open-Source Framework MOEA (Multiobjective Evolutionary Algorithms) bietet über eine abstrakte Schnittstelle mehr als 35 Evolutionäre Algorithmen an [216]. Es ist speziell hinsichtlich multikriterieller Optimierungsalgorithmen ausgerichtet. Zu den Algorithmen gehören unter anderem:

- Non-dominated Sorting Genetic Algorithm (NSGAII) [217]
- Multiobjective Cellular Genetic Algorithm (MOCell) [218]
- Pareto Archived Evolution Strategy (PAES) [219]
- Strength-based Evolutionary Algorithm (SPEA2) [220]
- Vector Evaluated Genetic Algorithm (VEGA) [221]

Zudem werden darüber die Optimierungsstrategien von den Libraries JMetal (Metaheuristic Algorithms in Java) [222] und PISA (Platform and Programming Language Independent Interface for Search Algorithms) [223] angebunden. Das Java basierte Framework steht unter der *GNU Lesser General Public License*.

5 Generisches Modell von zweistufigen Transportketten

Die Beschreibung des Konzeptes erfolgt gemäß der Aufteilung wie in der Abbildung 5.1. Zuerst wird der generische Modellierungsansatz über ein System- und Objektbeziehungsmodell beschrieben. Dazu sind die einzelnen Bestandteile flexibel anpassbar konzeptioniert und über einheitliche Schnittstellen miteinander verbunden. Anschließend erfolgt eine Erläuterung der Anwendbarkeit in Bezug auf die beiden betrachteten Domänen Logistik und Rechnernetze in den Abschnitten 5.2.1 und 5.2.2. Darüber hinaus wird der Einsatz im Zusammenhang mit einem Modell zur Geolokalisation in Abschnitt 5.2.3 beschrieben.

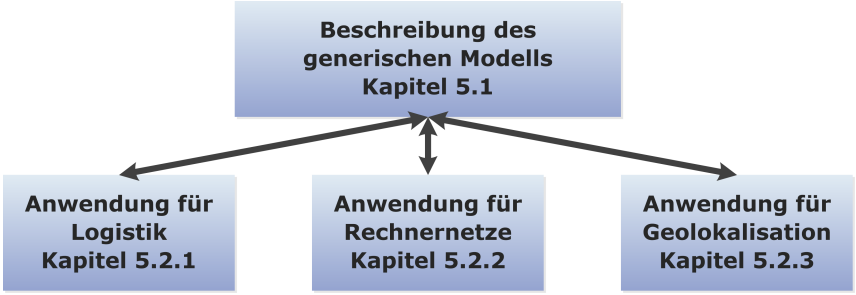


Abbildung 5.1: Vorgehensweise bei der Beschreibung des Modellierungsansatzes

5.1 Modellierung

Die Modellierung der generischen Transportkette sowie die Lösungsberechnung gemäß der Problemstellung in Kapitel 2 erfolgt in mehreren Prozessschritten. Das Systemmodell in der Abbildung 5.2 gibt dazu eine Übersicht der verschiedenen Teilprozesse und stellt das Vorgehen bei der optimierten Lösungsfindung dar. Zu Beginn wird von einem gegebenen Szenario mit dem generischen Ansatz ein Modell erstellt, welches zugehörige Komponenten und Transaktionen abbildet. Dieses enthält sämtliche Informationen, welche für die Bearbeitung in den weiteren Prozessen notwendig sind. Dazu gehören neben den verschiedenen Nebenbedingungen auch Zielfunktionen hinsichtlich der Optimierung und Simulation der Problemstellung. Zur Beschreibung der Prozesse wird davon ausgegangen, dass im Szenario noch keine Zentren festgelegt sind. Anderenfalls sind diese entsprechend im Modell mit einzupflegen. Anschließend wird für das entstandene Modell eine initiale, valide Lösung ermittelt. Dabei werden nur statische Optimierungskriterien berücksichtigt, da bis zu diesem Prozessschritt noch

keine Simulation erfolgt ist und diese eine valide Lösung voraussetzt. Danach erfolgt eine zyklische Optimierung der initialen Lösung hinsichtlich der definierten Zielkriterien. Diese sind in statische und dynamische Eigenschaften eingeteilt. Die Optimierung beinhaltet mehrere Teilprozesse. Die Reihenfolge der einzelnen Prozesse ist abhängig von dem verwendeten Optimierungsalgorithmus und lassen sich flexibel anpassen. Dies wird durch das gemeinsam genutzte Modell als Datenbasis ermöglicht. Zu den Teilprozessen gehört die Zuweisung der Nutzer zu den im Modell spezifizierten Zentren, die erneute Platzierung der Zentren und die Anpassung der Speicher. Nachdem diese drei Schritte abgearbeitet wurden, erfolgt die Simulation des Modells sowie eine Evaluation der Ergebnisse hinsichtlich der definierten Optimierungskriterien. Nach der Terminierung des Optimierungsverfahrens wird das Ergebnis in Form des Modells zurück gegeben. Dieses lässt sich auf Grund des einheitlichen Formats wieder mit einem anderen Optimierungsverfahren bearbeiten.

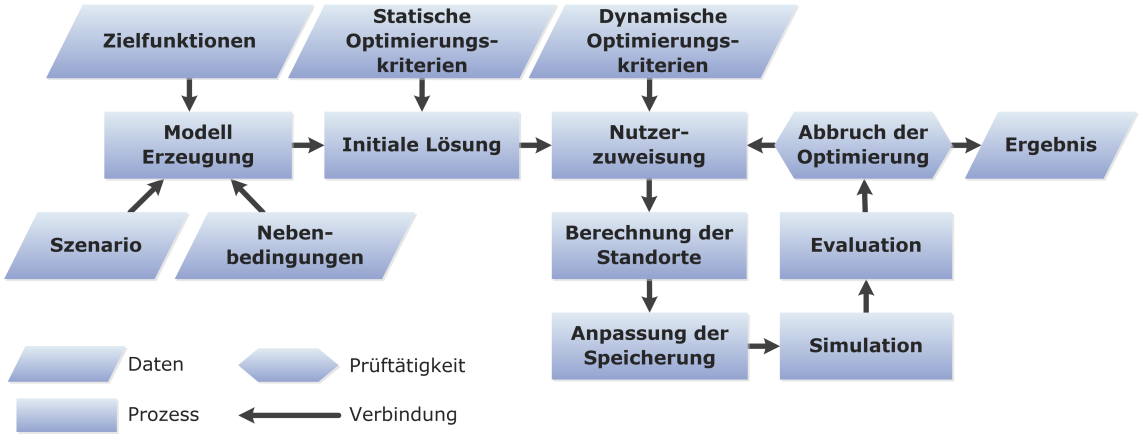


Abbildung 5.2: Programmstrukturplan der verschiedenen Teilprozesse des Systems für die Berechnung einer optimierten Lösung gemäß [224]

5.1.1 Abstraktionskonzept der Transportkette

Für die generische Modellierung von zweistufigen Transportketten werden die Problemstellungen von den verschiedenen Domänen mittels eines einheitlichen und generischen Ansatzes abgebildet. Dazu wird ein Modell erstellt, welches die Eigenschaften der verschiedenen Anwendungsgebiete abstrahiert und repräsentiert. Das Modell muss dabei unabhängig von dem betrachteten Anwendungsgebiet einer Transportkette sein. Es umfasst die gegebenen Informationen eines Szenarios und enthält Parameter für die Eigenschaften des zugehörigen Anwendungsgebietes. Ein Szenario stellt dabei die zu optimierende und simulierende Aufgabenstellung dar. Zur Beschreibung des Modells vom abstrahierten Szenario werden die definierten Begriffe von Kapitel 1.2.6 verwendet.

Zunächst erfolgt eine Analyse der Eigenschaften und Bestandteile von den verschiedenen Szenarien, bevor eine Modellierung mittels eines generischen Ansatzes erfolgt. Ein gegebenes Szenario besteht aus verschiedenen Kategorien von Eingangsdaten. Gemäß der Anforderungsanalyse in Kapitel 3 wird zwischen Fall-spezifischen, Domänen-

abhängigen und Domänen-unabhängigen Bestandteilen unterschieden. Dabei stellen die Domänen-unabhängigen Bestandteile die allgemeinen Eigenschaften dar, welche jedes entstehende Modell aufweisen muss. Die Domänen-abhängigen Merkmale sind über modulare und konfigurierbare Teilelemente in dem generischen Modell abzubilden. Diese lassen sich auch über Domänen-unabhängige Parameter mit einer neutralen Wertigkeit realisieren, wodurch die strukturelle Komplexität reduziert wird. Die Fall-spezifischen Informationen stellen die konkreten Werte der Parameter und Eigenschaften im Modell dar. Somit erfolgt die Beschreibung von abstrahierten Eigenschaften eines Szenarios auf Grundlage der Domänen-unabhängigen Bestandteile. Für die generische Anwendbarkeit werden diese entsprechend flexibel ausgestaltet.

Um mit einem einheitlichen und generischen Ansatz verschiedene Problemstellungen unterschiedlicher Domänen abzubilden, werden für die spezifischen Eigenschaften entsprechende verallgemeinerte Merkmale definiert. Dabei lassen sich die generellen Gemeinsamkeiten als Eigenschaften der Szenarien über Parameter und Schnittstellen abbilden. Ein Parameter oder eine Schnittstelle ist dabei so definiert, dass die Ausprägung der Eigenschaft eines Szenarios sich dadurch eindeutig beschreiben lässt. Durch die jeweilige Parametrierung wird sichergestellt, dass konkrete Ausprägungen einer Eigenschaft sich gegeneinander ausschließen und ein Szenario nur die spezifizierte Form annimmt. Verbleibende Unterschiede und spezifische Eigenschaften werden über modulare Teilelemente in dem Modell hinzugefügt bzw. über Parameter mit einer neutralen Ausprägung realisiert.

Zu den Domänen-unabhängigen Bestandteilen gehören die Eigenschaften, welche stets in einem Szenario der Transportkette unabhängig des Anwendungsgebietes vorhanden sind. Die betrachteten Aufgabenstellungen der Anwendungsgebiete FLP und CDN beruhen auf dem k-center und k-cluster Problem gemäß Kapitel 2. In Abhängigkeit des spezifischen Szenarios sind dabei die verschiedenen Varianten des Problems mit Parametern zu berücksichtigen. Somit sind dessen Elemente feste Bestandteile des Modells. Dazu gehört die Menge an Nutzern und Zentren sowie die Infrastruktur. Weiterhin gibt es gemäß der Problembeschreibung in Kapitel 2 in jeder Transportkette die Dienste, welche zwischen den Teilnehmern transferiert werden. Für die Lösung der Problemstellung und Optimierung des Ergebnisses besitzt das Modell eine Bewertungs- und Evaluierungsfunktion hinsichtlich einer oder mehrerer Zielkriterien. Die Domänen-unabhängigen Bestandteile werden im Folgenden kurz in deren Funktion erläutert, bevor diese im Detail modelliert werden. Die Abbildung 5.3 gibt dazu eine Übersicht der im Modell vorhandenen Entitäten sowie deren wesentlichen Eigenschaften.

Die einzelnen Eingangsdaten werden im Folgenden stückweise in das generische Modell überführt. Dazu wird für jeden Bestandteil entsprechend seiner Abstraktion eine Modellierung vorgenommen.

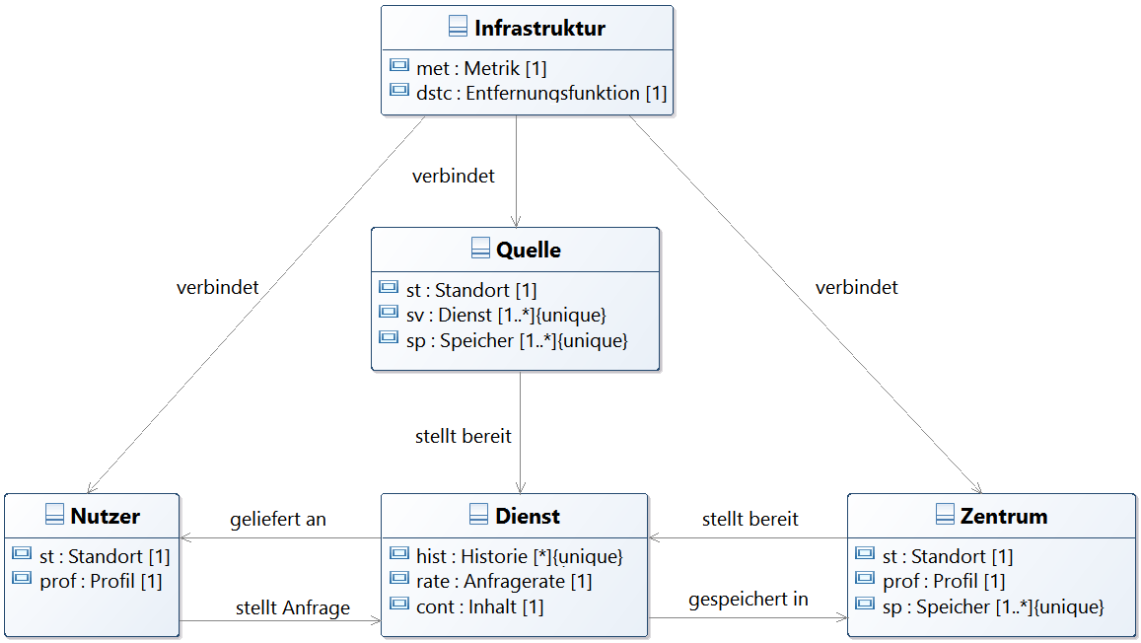


Abbildung 5.3: Bestandteile des abstrahierten Szenarios einer Transportkette gemäß UML Klassendiagramm

Nutzer

Unter dem Begriff Nutzer wird diejenige Entität verstanden, welche Anfragen an das System des Transportnetzes stellt und die zur Verfügung gestellten Dienste anfordert. Jeder Nutzer hat dabei mehrere Eigenschaften. Ein Nutzer besitzt einen spezifischen Standort, von welchem aus die Anfrage gestellt wird und wo die Antwort hin zu liefern ist. Darüber hinaus verhält sich ein Nutzer entsprechend einem individuellen Interessen-Profil, nachdem dieser Anfragen stellt. Dadurch werden spezifische Dienste in unterschiedlicher Intensität von einem Zentrum angefordert. Das individuelle Verhalten ist dementsprechend gemäß dem definierten Profil eines Nutzers zu modellieren. Auf Grund der Vielfältigkeit und Funktionalität hinsichtlich der Simulation, die ein Profil annehmen kann, wird dieses zur Übersichtlichkeit als eigenständiges Teilelement des Nutzers modelliert.

Zentrum

Die Bedeutung eines Zentrums entspricht der Rolle einer Zwischenstation in der Transportkette. Es fordert Dienste von einer Quelle an, speichert diese im zugehörigen Speicher zwischen und gibt auf Anfrage die Dienste wieder ab. Dafür sendet ein Nutzer seine Anfrage an das System und ein Zentrum liefert den angeforderten Dienst an den spezifizierten Ort. Somit übernimmt ein Zentrum die Rolle der Verteilung von Diensten. Den Zentren ist die spezifische Quelle für einen Dienst bekannt. Ein Zentrum muss zur Bereitstellung eines angeforderten Dienstes diesen zuvor im eigenen Speicher haben. Daher ist die Verfügbarkeit eines Dienstes bei dem angefragten Zentrum stets zu berücksichtigen. Da der Speicher unterschiedliche Verhaltensweisen bei der Verwaltung der verfügbaren Kapazität und hinsichtlich der Simulation annehmen kann, wird dieser

als eigenständiges Teilelement des Zentrums modelliert. Die über die Laufzeit hinweg angeforderten Dienste an einem Zentrum stellen ebenfalls eine Art Profil auf, welches vergleichbar mit dem Profil der Nutzer ist. Somit lässt sich auch für ein Zentrum aus den Nutzeranfragen und den verfügbaren Diensten im Speicher ein Profil ermitteln.

Quelle

Die Quelle dient der initialen Bereitstellung von Diensten einer vordefinierten Art. Diese stellt selbst eine Art Zentrum dar, welches keine Zwischenspeicherung von weiteren Diensten vornimmt. Eine Quelle beliefert ein Zentrum oder auch einen Nutzer auf Anfrage.

Dienst

Der Dienst stellt die Abstraktion der von einem Transportsystem bereitgestellten und übertragenen Verkehrsobjekte dar. Es ist eine ganzheitliche Entität, welche alles das zusammenfasst, um die Anforderungen eines Nutzers hinsichtlich eines spezifischen Zwecks zu erfüllen. Dieser ist dabei in Abhängigkeit des Anwendungsgebietes ein körperlicher oder ein immaterieller Gegenstand. In der Logistik zählen zu den körperlichen Gegenständen unter anderem Produkte, Güter und Waren. Bei Rechnernetzen stellen Daten die immateriellen Gegenstände dar. Zudem ist es in beiden Bereichen unerheblich, ob diese direkt zur Verfügung stehen oder dynamisch als Dienstleistung erbracht werden Ein Dienst wird über die Infrastruktur von der Quelle bis hin zum Nutzer transportiert. Dienste unterscheiden sich in deren Ausprägung maßgeblich durch Art, Typ und Gestalt.

Infrastruktur

Der Transport der Dienste erfolgt über eine Infrastruktur, welche somit die Grundlage des Systems bildet. Dieser Bestandteil verbindet die Entitäten Nutzer, Zentrum und Quellen miteinander. Über die Infrastruktur wird die Entfernung zwischen den Teilnehmern gemessen und auch die Metrik für die Entfernung festgelegt. Es abstrahiert die Verbindungsstrecken des jeweiligen Anwendungsgebietes über eine spezifische Entfernungsberechnung. Auf Grund unterschiedlicher Möglichkeiten in der Berechnung und den definierten Metriken erfolgt diese über eine separate Funktion. Diese verwendet eine einheitliche Schnittstelle zur Infrastruktur und bildet somit einen separaten Bestandteil des Modells.

5.1.2 Modellierung des Nutzers

Die Abstraktion zum Nutzer gemäß der Definition aus Kapitel 1.2.6 stellt bereits den ersten Schritt für den generischen Ansatz dar. Die in einem Szenario gegebenen Nutzer werden als eigenständige Entitäten im Modell abgebildet, da diese in der Ausprägung der Eigenschaften unterschiedlich sind. Zur Differenzierung besitzt jeder Nutzer eine eindeutige Bezeichnung, die über eine Identifikator ID gegeben ist. Dieser wird bei der

Initialisierung vergeben und dient der direkten Adressierung. Weiterhin hat ein gegebener Nutzer einen geographischen Standort. Dafür erhält dieser die Eigenschaft, welche ebenfalls initial zu definieren ist. Somit sind jedem Nutzer bei der Modellerstellung seine Koordinaten zuzuweisen. Die Angabe erfolgt flexibel als mehrdimensionaler Punkt. Dieser wird in den betrachteten Anwendungsgebieten durch die Parameter Latitude (Breitengrad) und Longitude (Längengrad) beschrieben. Die Information ist gemäß der definierten Problemstellung aus Kapitel 2 als Eingangsdaten des zugrundeliegenden Szenarios vorgegeben. Zur Reduktion der Komplexität lassen sich mehrere, dicht beieinander liegende Nutzer zu einer Entität zusammen fassen und deren weitere Eigenschaften aggregieren. Dies ist insbesondere möglich, wenn eine Gruppe von Nutzern den gleichen Zugangspunkt zur Infrastruktur besitzen. Weiterhin erhält der Nutzer eine Priorität zugewiesen. Dieser Wert lässt sich flexibel durch die Eingangsdaten definieren. Hierdurch wird die Möglichkeit geschaffen, eine Unterscheidung zwischen den Nutzern hinsichtlich der Wichtigkeit vorzunehmen. Zum Beispiel wichtet der Dienstanbieter damit verschiedene Nutzergruppen. Der Standardwert ist mit eins als neutrales Element spezifiziert und lässt sich in Abhängigkeit weiterer Eingangsdaten bestimmen. Ein höherer Wert steigert dessen Priorität und somit dessen Wertigkeit proportional.

Zudem werden bei dem Nutzer die Informationen zu den Zentren vorgehalten, insbesondere zu dem primär zugewiesenen Zentrum. Diese beinhalten unter anderem die Entfernung zwischen dem Nutzer und dem jeweiligen Zentrum sowie einen Korrelationswert hinsichtlich der Interessen. Darüber hinaus hat der Nutzer die Eigenschaft eines Interessen-Profiles. Dieses gibt an, welche Anfragen an das System des Transportnetzes gestellt werden. Auf Grund der Komplexität und der mehrfachen Verwendung wird das zugehörige Teilmodell in Abschnitt 5.1.4 beschrieben.

5.1.3 Modellierung des Dienstes

Zur erweiterten Anwendbarkeit von Transportketten in einem weiter gefassten Kontext werden die beförderten Verkehrsobjekte der Kette im generischen Modell als Dienst abstrahiert. Dieser ist dabei als real oder virtuell anzusehen, um flexibel in Abhängigkeit des betrachteten Anwendungsbereiches die Verkehrsobjekte abzubilden. Verschiedene Dienste werden im Modell maßgeblich durch die Eigenschaft der Bezeichnung voneinander unterschieden. Ein Dienst besitzt im Modell die Eigenschaft der Menge, welche stets mit zu spezifizieren ist. Der Wertebereich nimmt entweder einen diskreten Wert an, sodass sich Stückgüter abbilden lassen, oder stellt ein kontinuierliches Angebot dar, um das Replizieren von digitalen Daten zu modellieren. Weiterhin weist ein Dienst ein definiertes Format auf. Dieses umfasst diejenigen beschreibenden Eigenschaftsmerkmale, welche in Abhängigkeit der Problemstellung und der zu analysierenden Aspekte für die Optimierung und Simulation eine Relevanz haben. Eine Teileigenschaft des Formats ist dabei stets die Größe, welche die zu belegende Menge des Speicherplatzes in einem Zentrum angibt.

5.1.4 Modellierung des Interessen-Profils

Gemäß den Anforderungen aus Kapitel 3 ist im generischen Modell vorgesehen, dass ein Nutzer ein spezifisches Interessen-Profil aufweist. Dazu wird die bisherige Modellierung eines Nutzers um die Eigenschaft des Profiles ergänzt. Dieses lässt sich als eigenständiges Teilmodell ansehen und wird daher hier separat beschrieben.

Das Interesse eines Nutzers wird durch mehrere Eigenschaften charakterisiert. Wie auch in der Realität haben Nutzer Interesse an unterschiedlichen Diensten. Somit ist die Art des Dienstes als Merkmal zu unterscheiden. Weiterhin besteht eine Variation in der Menge der Dienste, die im Interessenbereich eines Nutzers liegen. Das Profil wird daher auch über die Anzahl der durch einen Nutzer angefragten Dienste bestimmt. Als weiteres Merkmal ist die Intensität zu unterscheiden, mit welcher ein Dienst angefragt wird. Darüber hinaus bevorzugt jeder Nutzer einen Dienst mit einer anderen Priorität und Wahrscheinlichkeit.

Zur Modellierung der Interessen eines Nutzers werden die zuvor genannten Eigenschaften als Profil umgesetzt. Das Modell des Profils wird durch die Anfragehäufigkeit beschrieben. Diese gibt an, mit welcher Wahrscheinlichkeit ein Nutzer eine Anfrage an das System aus Zentren stellt, wenn dieser betrachtet wird. Der Wert ist zwischen 0 %, Nutzer stellt nie eine Anfrage, und 100 %, Nutzer stellt immer eine Anfrage zum nächst möglichen Zeitpunkt, zu spezifizieren. Die Anfrageverteilung von Diensten wird über eine Tabelle abgebildet. Diese enthält für jeden Dienst im Interessenbereich des Nutzers einen Eintrag mit dessen Bezeichnung und der Anfrage-Wahrscheinlichkeit sowie Intensität. Zudem werden im Profil vergangene Anfragen in sogenannten Traces gespeichert. Diese lassen sich auch als Eingangsdaten des Szenarios vorgeben und in der Optimierung und Simulation nutzen. Zum einen lassen sich Vorhersagen treffen und zum anderen ermöglicht dies eine Wiederholung der Simulation mit exakt den gleichen Daten.

Da sich die Profile von Nutzern über die Zeit ändern, lässt sich für die Anfragehäufigkeit und für jeden Eintrag in der Tabelle der Anfrageverteilung eine zeitabhängige Funktion hinterlegen, welche den jeweiligen Wert zu einem Zeitpunkt bestimmt. Dadurch wird ein sich dynamisch änderndes Profil abgebildet, welches einen Wechsel der Interessen ermöglicht und Einfluss auf aktuelle Geschehnisse wiedergibt.

Die Tabelle 5.1 gibt ein Beispiel für die Modellierung von Interessen-Profilen.

Tabelle 5.1: Beispiel einer Anfrageverteilung von einem Nutzer

Dienst	Wahrscheinlichkeit (ω)	$\omega(t)$	Intensität (*int*)	*int*(t)
A	50 %	$0{,}5 + 0{,}1 \cdot \sin(t)$	1	konstant
B	30 %	$0{,}3 - 0{,}1 \cdot \sin(t)$	2	konstant
C	10 %	konstant	3	konstant
D	10 %	konstant	1	$1+(-1)^t$

In Bezug zum Modell des Gesamtsystems enthält das generische Modell neben der Zielfunktion eine Schnittstelle für eine Korrelationsfunktion. Über diese konfigurierbare Funktion lassen sich gleichartige und ähnliche Interessens-Profile identifizieren und in die

Optimierung mit einbeziehen. Darüber hinaus wirken sich die Profile der Nutzer direkt auf das Systemverhalten während der Simulation aus.

Zur Verwendung von Nutzerprofilen sind zwei Verfahren zu unterscheiden. Zum einen werden authentische Daten genutzt, indem aufgezeichnete Traces oder Umfragen als Eingangsdaten für ein Szenario zur Verfügung stehen. Zum anderen besteht die Möglichkeit diese Daten synthetisch zu generieren. Der Vorteil von authentischen Daten besteht in dem direkten Zusammenhang von Modell und Realität. Allerdings ist bei deren Nutzung der Datenschutz zu beachten. Um diesem Problem zu entgehen, lässt sich eine synthetische Methode auf Basis eines validen Modells nutzen, um realistische Profile zu erzeugen.

Interessen-Profil auf Grundlage authentischer Informationen

Die Interessen eines Nutzers lassen sich über direkte Befragungen und Marktanalysen ermitteln. Dazu wird die Zielgruppe hinsichtlich der durch das System angebotenen Dienste analysiert und gemäß der Parameter von der Modellierung ein Profil erstellt. Eine weitere Möglichkeit das Interessen-Profil zu erhalten, besteht in der Aufzeichnung des Nutzerverhaltens. Hierbei wird das Profil indirekt ermittelt. Dazu werden die Anfragen eines Nutzers an das System in der betrachteten Problemstellung über einen Zeitraum gespeichert und als Trace vorgehalten. Aus dieser Historie von Nutzeranfragen lassen sich Anfragewahrscheinlichkeiten für einen Dienst ableiten. Durch Häufigkeitsanalyse wird aus einem Trace rückwirkend das Profil eines Nutzers ermittelt und die Anfrageverteilung des Profiles berechnet. Über eine Einteilung des Traces in Zeitfenster lassen sich zwischen den einzelnen Abschnitten Unterschiede aufzeigen und somit Tendenzen in der Veränderung bestimmen, welche sich wiederum in der Profilerstellung auswirken.

Die Tabelle 5.2 zeigt den aufgezeichneten Trace *multi2* [225]. Die Folge stellt dabei die zeitliche Reihenfolge der Anfragen dar und die angeforderten Dienste sind mit einer Nummer abstrahiert. Die Abbildung 5.4 stellt die ersten 5000 aufgezeichneten Einträge des Traces *sprite* [226] dar. Beide Traces beziehen sich auf die angeforderten Datenblöcke einer Festplatte.

Tabelle 5.2: Beispiel eines Traces anhand der ersten, aufgezeichneten Referenzen des Datensatzes *multi2* [225]

Folge	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Dienst	0	1	2	3	4	5	6	7	7	8	8	9	10	11	12	13	14	15	16	0	1	2	17	1	18

Interessen-Profil mittels parametrieter Generierung

Ein Szenario lässt sich auch mittels dem Einsatz von synthetischen Interessen-Profilen simulieren und analysieren. Dazu werden dem generischen Modell Daten übergeben, welche sich auf Basis eines validen Modells mit variablen Parametern generieren lassen. Dies bietet eine hohe Flexibilität für Experimente und Analysen. Dazu hat sich in der Wissenschaft das Modell der Zipf-Verteilung durchgesetzt [145, 146, 165]. Dieses stammt

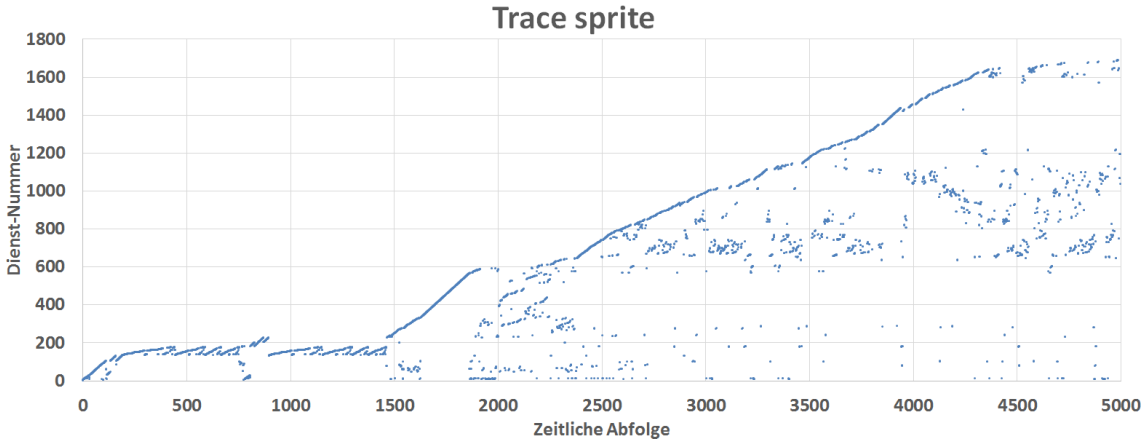


Abbildung 5.4: Die ersten 5000 aufgezeichneten Einträge des Traces *sprite* [226] von über 130.000

aus der Linguistik und beschreibt die Häufigkeit des Auftretens eines Wortes bei einem hinreichend langen Text [227]. Es entspricht dem Inversen proportional zu dessen Rang. Die Anfragewahrscheinlichkeit eines Dienstes ergibt sich somit gemäß seinem $Rang^{-\alpha}$ des Interessen-Profils [228]. Mit steigendem Wert für α ergeben sich Profile mit ungleichmäßigerer Verteilung in der Anfragewahrscheinlichkeit. Der Parameter α ist spezifisch in Abhängigkeit der Domäne. Die Abbildung 5.5 zeigt den Vergleich zwischen der Wortverteilung in der Literatur von Goethes Faust und der Zipf-Verteilung. Diese Verteilungskurve spiegelt sich auch bei weiteren Texten wieder [229].

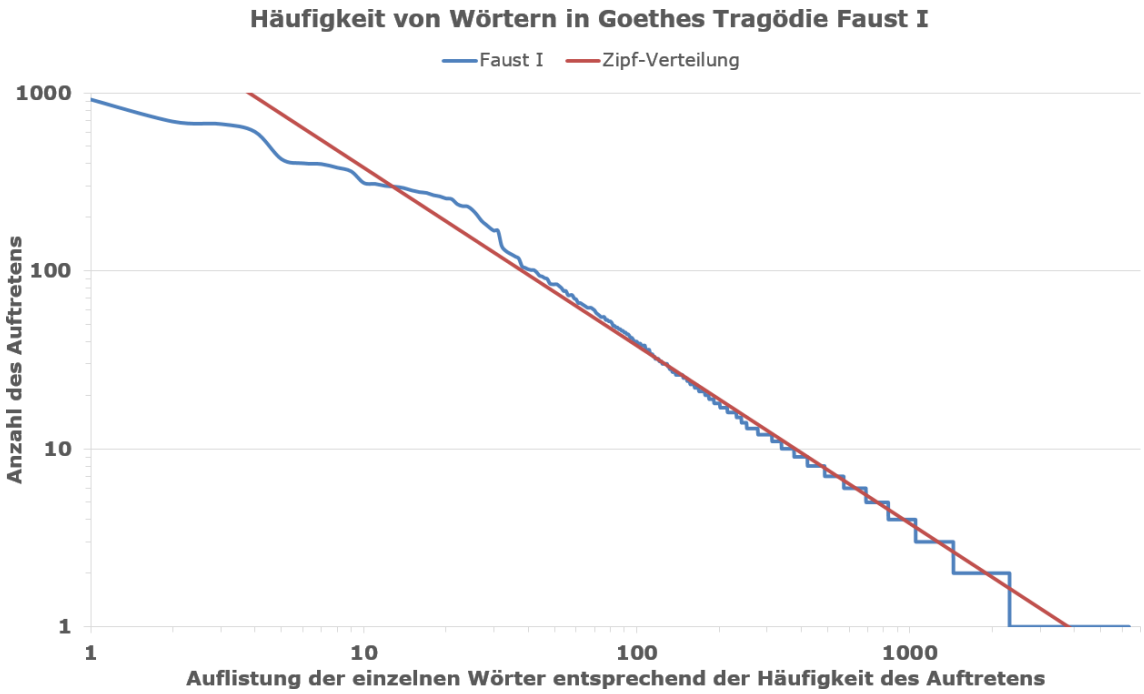


Abbildung 5.5: Verteilung von Wörtern in Goethes Werk Faust I [230] im Vergleich zur perfekten Zipf-Verteilung mit $\alpha = 1$

Im Bezug auf die betrachtete Problemstellung folgt die Interessensverteilung eines Nutzers ebenfalls der Zipf-Verteilung. Dies wurde in mehreren Analysen in der wissenschaftlichen

Literatur nachgewiesen, sodass sich mit diesem Modell Profile von Nutzern nachbilden lassen [149, 155, 231]. Die Abbildung 5.6 zeigt den Vergleich von realen Daten und der Zipf-Verteilung im Bereich von Webseiten sowie den damit verbundenen Datenmengen.

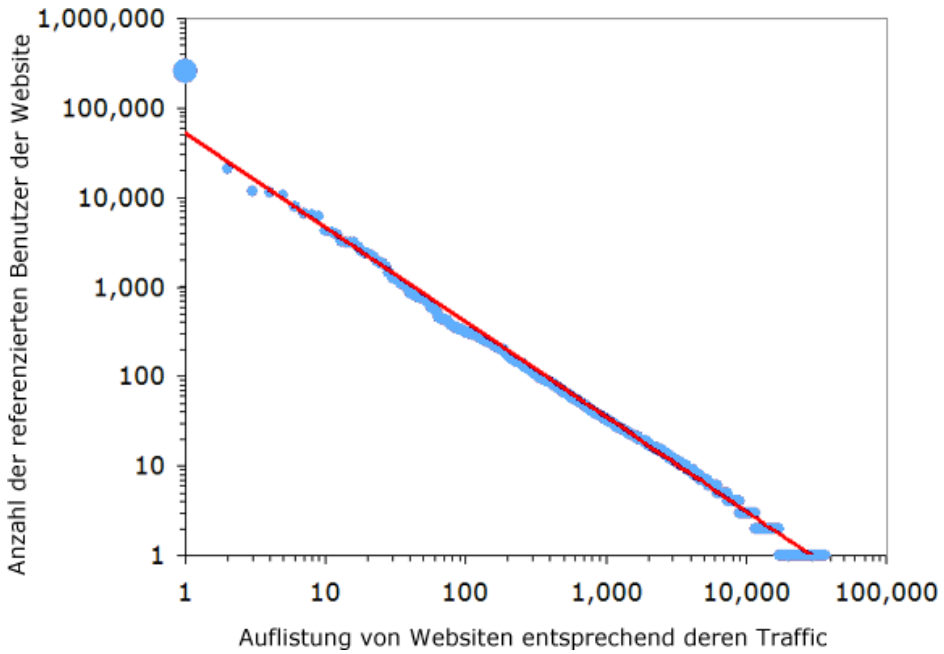


Abbildung 5.6: Vergleich von realen Daten und der Zipf-Verteilung bei referenzierten Webseiten entsprechend deren Datenverkehr und der Anzahl der Besucher (Vgl. [155, 231, 232])

Bei der Abbildung von Profilen kann es insbesondere bei sehr kleinen und sehr großen Rängen zu unzureichenden Genauigkeiten kommen. Für eine präzisere Beschreibung bietet sich die Erweiterung mit $(Rang + c_1)^{-(\alpha+c_2)}$ an [233]. Der Parameter c_1 ermöglicht eine Anpassung im Bereich der niederen Ränge und der Parameter c_2 gewährt eine Anpassung im Bereich der hohen Ränge.

5.1.5 Modellierung des Zentrums

Die Zwischenstationen in einer Transportkette werden als Zentren abstrahiert. Diese übernehmen die Rolle der Zwischenspeicherung von Diensten sowie deren Verteilung zu den Nutzern auf Anfrage. Dabei wird durch das Szenario spezifiziert, wie viele Zentren das zugehörige Modell maximal besitzen darf. Mit der computergestützten Optimierung und Simulation lassen sich mehrere Werte für ein Szenario analysieren.

Ein Zentrum hat mehrere Eigenschaften und Funktionen, damit es die zugewiesene Rolle hinreichend übernimmt. Ein Zentrum ist eindeutig durch seine ID identifizierbar. Es besitzt die Eigenschaft eines Standortes. Dieser wird wie bei den Nutzern über die Koordinaten eines mehrdimensionalen Punktes beschrieben. Die Festlegung der Koordinaten erfolgt während der Initialisierung und wird in der Optimierung angepasst. Durch den Standort ist die Anbindung zu der gegebenenfalls vorgegebenen Infrastruktur definiert.

Ein Zentrum kennt alle seine zugewiesenen Nutzer. Darüber hinaus besitzt es ein Profil, obwohl es selbst kein eigenes Verhaltensmuster wie die Nutzer aufweist. Dieses repräsentiert die am jeweiligen Zentrum ankommenden Anfragen der Nutzer und somit die voraussichtlich angebotenen Dienste. Das Profil des Zentrums ergibt sich aus den Profilen der primär dem Zentrum zugewiesenen Nutzer.

Für die Zwischenspeicherung von Diensten besitzt das Zentrum einen Speicher. Die Größe des Speichers wird über die maximale Kapazität initial festgelegt.

Mit dem Speicher wird auch die Verwaltungsstrategie definiert. Hierbei sind zwei grundsätzliche Prinzipien zu unterscheiden. Diese wirken sich insbesondere auf das Systemverhalten in der Simulation aus. In Abhängigkeit der Art des Dienstes erfolgt entweder eine persistente oder nicht persistente Speicherung.

Bei der persistenten Speicherung erhält ein Zentrum die Dienste mit einer entsprechenden Anzahl zugesendet. Diese Mengen von gleichartigen Diensten werden im Zentrum vorgehalten und belegen entsprechend der Anzahl die Kapazität im Speicher. Hierbei muss ein Dienst im Zentrum solange gespeichert werden, bis dieser aufgebraucht ist. Dies ist dann der Fall, wenn die vorgehaltene Anzahl auf Null sinkt. Dem Zentrum ist es nicht gestattet Dienste auf Grund von Mangel an Kapazität zu entfernen. Die Dienste sind durch das Anfordern der Nutzer zu konsumieren. Für die Verwaltung des kapazitätsbegrenzten Zwischenspeichers besitzen die Zentren eine Strategie zur Anforderung von Diensten mit entsprechender Kapazität. Dieses Prinzip der persistenten Speicherung findet vor allem bei physischen Diensten statt, welche sich nicht einfach duplizieren lassen.

Bei der nicht persistenten Speicherung bewahrt das Zentrum den zugesendeten Dienst ebenfalls in seinem Speicher auf. Hierbei ist es jedoch dem Zentrum freigestellt einen Dienst im Zwischenspeicher selbständig bei Bedarf zu entfernen, unabhängig von der verbleibenden Anzahl. Somit müssen die Dienste nicht von den Nutzern durch entsprechende Anforderungen aufgebraucht werden. Dies erhöht die Flexibilität bei der Verwaltung des Zwischenspeichers durch das Zentrum, insbesondere bei ausgeschöpfter Kapazität des Speichers. Erhält das Zentrum in solch einer Situation eine Anfrage nach einem gegenwärtigen nicht im Speicher verfügbaren Dienst, dann ist dieses dennoch in der Lage die Anfrage zu erfüllen. Hierbei wählt das Zentrum einen Dienst in seinem Zwischenspeicher aus und entfernt diesen, um freie Kapazität zu schaffen. Anschließend fordert das Zentrum selbst den vom Nutzer gewünschten Dienst bei einer anderen Quelle an und speichert diesen zwischen. Danach wird die Anfrage des Nutzers positiv mit dem Dienst erfüllt. Zur Verwaltung des Zwischenspeichers werden entsprechende Ersetzungsstrategien angewendet. Die Anzahl eines vorgehaltenen Dienstes ist hierbei von nachrangiger Bedeutung. Dieses Prinzip der nicht persistenten Speicherung erfolgt vorrangig bei digitalen Diensten, welche sich duplizieren lassen und keine wirtschaftlichen Nachteile durch das Entfernen aus dem Speicher entstehen.

5.1.6 Modellierung der Infrastruktur

Unabhängig der verschiedenen Anwendungsgebiete werden die Verbindungsstrecken zwischen den Nutzern durch die Infrastruktur abstrahiert und zusammengefasst. Hierbei

ist es freigestellt, ob es sich um eine real gegebene Struktur oder ein virtuell spezifiziertes Overlay-Netz handelt. Das Overlay-Netz entspricht einem logischen Netz basierend auf einer real existierenden Struktur. Als Beispiel einer solchen Klasse von Netzen sind *Virtual Network Embedding* zu nennen [234], welche im Rahmen von CDN auftreten. Beim FLP wird vorrangig das Straßennetz genutzt, wohingegen bei CDN Rechnernetze die Nutzer verbinden. Die Infrastruktur dient zur Übertragung von Diensten in der jeweiligen Problemstellung. Diese stellt die Grundlage der Entfernungsermittlung dar und ist in Zusammenhang mit der Funktion zur Entfernungsberechnung zu betrachten. Insbesondere gibt es verschiedene Metriken, welche durch die jeweilige Ausprägung der abstrahierten Entfernungsfunktion zu berechnen sind. Alle Nutzer des Szenarios müssen an die definierte Infrastruktur angeschlossen sein, damit diese auch die angefragten Dienste erhalten. Das gesamte System ist direkt von der zugrundeliegenden Infrastruktur abhängig.

Abstraktion der Infrastruktur

Die Infrastruktur wird formal als ein Graph mit Knoten und Kanten aufgefasst, wodurch es sich direkt im Modell abbilden lässt. Bei den Knoten ist zwischen Nutzer-Knoten und Verbindungs-Knoten zu unterscheiden. Nutzer-Knoten repräsentieren die Nutzer mit deren geographischen Standort. Verbindungs-Knoten bilden die Schnittpunkte von mehreren Verbindungen und dienen zur direkten Modellierung der Infrastruktur. Diese besitzen im Gegensatz zu den Nutzer-Knoten keine weiteren Eigenschaften. Die Kanten entsprechen der Verbindung der einzelnen Knoten untereinander. Hierbei ist davon auszugehen, dass keine regelmäßig strukturierte Topologieform vorliegt. Der jeweilige Nutzer mit seinem Standort wird durch einen Knoten repräsentiert, welcher den gleichen geographischen Standort besitzt. Sofern Verbindungen durch das Szenario vorgegeben sind, werden diese in das generische Modell übernommen und berücksichtigt. Anderenfalls ist von einem vollständig vermaschten Netz auszugehen. Dabei besitzt jeder Knoten eine direkte Verbindung zu jedem anderen Knoten. Die Länge der Verbindungsstrecke wird durch die Entfernungsfunktion berechnet. Eine Verbindung besitzt mehrere Merkmale, welche im Modell durch die entsprechende Kante mit deren Eigenschaften geeignet zu repräsentieren ist. Die wesentlichen Unterschiede bestehen in der Länge, der Verbindungskapazität, der Verzögerungszeit sowie der Auslastung. In Abhängigkeit der Metrik und des Zielkriteriums während der Berechnung einer optimierten Lösung werden diese Verbindungseigenschaften zu einer virtuellen Länge aggregiert. Dieser Wert dient zur genaueren Differenzierung zwischen verschiedenen Kanten und wird als das Gewicht einer Kante bezeichnet.

Abstraktion der Entfernungsfunktion

In den verschiedenen Anwendungsgebieten werden unterschiedliche Metriken bei der Berechnung einer Verbindungsstrecke als auch bei der Verbindungswahl genutzt. Im Bereich des FLP sind die Metriken Länge in Kilometer und notwendige Zeit einer Verbindung von Bedeutung. Demgegenüber werden im Bereich von CDN die

Längen vorrangig in Hops angegeben [235]. Darüber hinaus sind die Eigenschaften Bandbreite, Auslastung und Latenz von Belang. Daher erfolgt die Abstraktion zu einer Entfernungsfunktion mit einer einheitlichen Schnittstelle. Diese ist Bestandteil des generischen Modells eines Szenarios und ist vorab zu spezifizieren. Hierbei gibt es mehrere Ausprägungen, welche die Entfernungsfunktion in Abhängigkeit der Konfiguration annimmt. Im Folgenden sind mehrere Möglichkeiten der Ausprägung gegeben:

- Euklidisch mit $d(x,y) = \|x-y\|_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Manhattan mit $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$ [236]
- Orthodrome (gemäß *World Geodetic System 1984* - Referenz Ellipsoid [237])
- Kartenbasiert durch Nutzung von Kartendiensten (z. B. Google Maps [238])
- Graphbasiert mit Berechnung kürzester Pfade
  (z. B. Algorithmus von Dijkstra [239], Floyd-Warshall [240, 241] oder A*-Algorithmus [242])
- Tourbasiert zur Verbindung mehrerer Knoten auf einer Strecke
  (z. B. nach Handlungsreisenden [243])

Wenn eine Infrastruktur auf Grund der hohen Komplexität, der Verfügbarkeit oder der Menge an Daten nicht explizit vorgegeben wird, lässt sich diese dennoch implizit im generischen Modell nutzen. Dazu wird die Infrastruktur über eine Schnittstelle an das Modell angeschlossen. Dabei übernimmt die Entfernungsfunktion die Rolle eines Schnittstellen-Adapters und übermittelt die Anfragen an den jeweiligen Anbieter. Als Beispiel ist hier die Anbindung der Kartendienste Google Maps [238] oder OpenStreetMap [244] zur Entfernungsberechnung zwischen zwei geographischen Standorten zu nennen. Nach diesem Ansatz ist auch bei der komplexen Problemstellung einer Tourenplanung zu verfahren.

5.1.7 Abstrahiertes System

Die Abbildung 5.7 fasst die wesentlichen Bestandteile des abstrahierten Szenarios zusammen und gibt einen Überblick vom generischen Modell. Die Knotenpunkte repräsentieren die Nutzer mit deren Standort in Relation zueinander. Die zugehörigen Prioritäten sind mit angegeben. Zur Reduktion der Komplexität lässt sich auch eine Gruppe von Nutzern auf einen Knoten abbilden, indem diese in den Eigenschaften aggregiert werden und ein gemeinsamer Standort spezifiziert wird. Jedem Knoten ist ein Profil mit dem Kreisdiagramm zugewiesen, wobei die anteilige Größe der markierten Kreisfläche das durchschnittliche Interesse an einem Dienst angibt. Bei einer Zusammenfassung von Nutzern erfolgt dazu eine Aggregation der einzelnen Profile von den entsprechenden Gruppenteilnehmern. Die Kanten zwischen den Knoten zeigen die Infrastruktur und die Verbindungen der Knoten untereinander. Die Eigenschaften einer Kante werden über das

jeweilige Gewicht angegeben. Die Standorte der Zentren sind durch die rot markierten Knoten angezeigt. Das dabei angegebene Kreisdiagramm ist das durchschnittliche Profil von den zugewiesenen Nutzern und stellt somit die an dem jeweiligen Zentrum eingehenden Anfragen nach Diensten dar. Die Speicherkapazität eines Zentrums erfolgt durch eine Angabe am Knoten. Auf Grund der Übersichtlichkeit ist die Zuweisung der Nutzer zu deren primären Zentrum nicht mit dargestellt, zumal diese sich während der Simulation und Optimierung ändert.

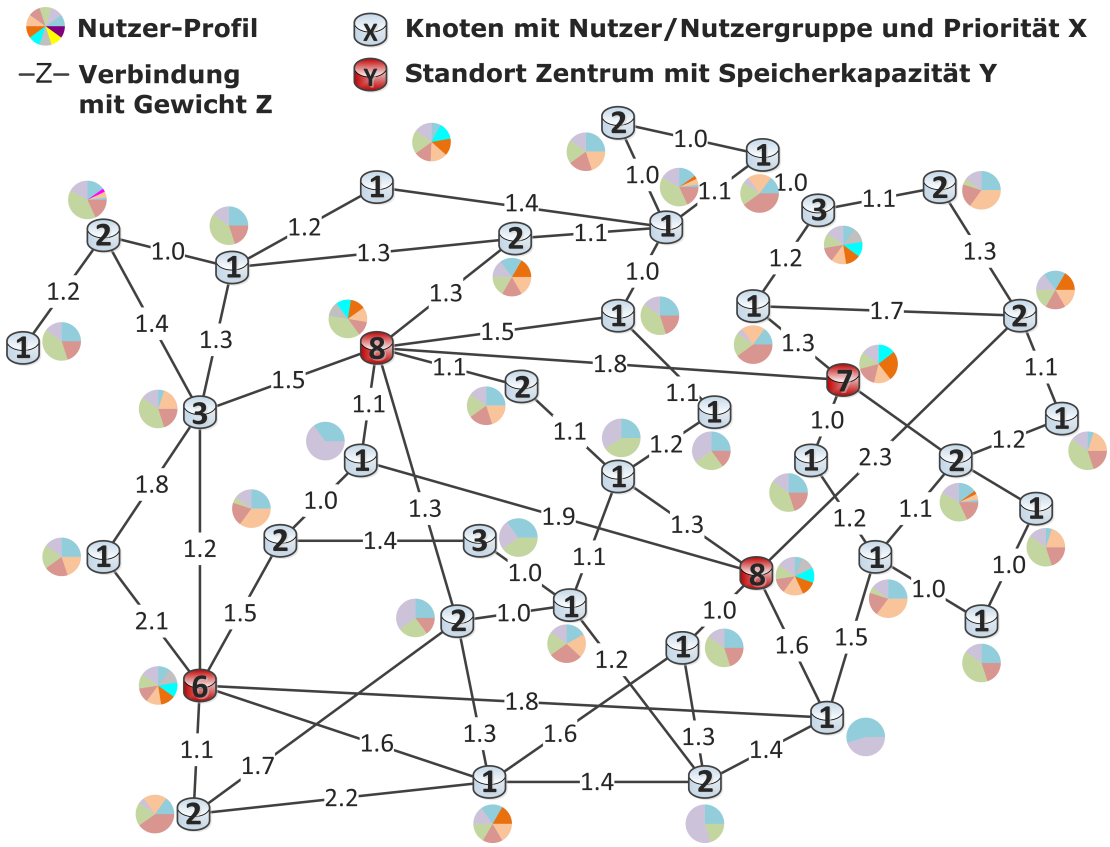


Abbildung 5.7: Generisches Modell des abstrahierten Szenarios einer zweistufigen Transportkette zur Planung und Optimierung von Zentren

5.1.8 Einbeziehung von Nebenbedingungen

In Abhängigkeit der Problemstellung aus Kapitel 2 werden bei der Modellerzeugung und optimierten Lösungsfindung Randbedingungen vorgegeben. Einer der wesentlichen Aspekte bei der Optimierung ist gemäß Kapitel 2.3 die Anzahl der Zentren für ein Szenario. Für die Analyse der notwendigen Anzahl ist im Modell ein Parameter spezifiziert, welches die maximale Menge der Zentren für ein Szenario vorgibt. Auf Grund der NP-Schwere des Problems ergeben sich bei unterschiedlichen Vorgaben andere Aufgabenstellungen, sodass diese unabhängig voneinander zu optimieren sind. Daher wird für diesen Optimierungsaspekt im generischen Modell ein Parameter spezifiziert, welcher vor der Lösungsberechnung zu definieren ist. Die Modellierung gemäß dem vorgestellten Ansatz bleibt dabei unverändert. Über die Variation mit verschiedenen

Werten des Parameters lassen sich für ein Szenario mehrere Fälle in der Optimierung und Simulation entsprechend analysieren.

Eine weitere Nebenbedingung ist die Platzierungsmöglichkeit von Zentren. In Anbetracht der unterschiedlichen Anwendungsgebiete ergeben sich die beiden Extreme der knotenbasierten Platzierung und der freien Platzierung, welche bei der Lösungsfindung zu beachten sind. Weiterhin bestehen verschiedene Mischformen von den Bedingungen, wie z. B. das eine freie Platzierung nur innerhalb bestimmter Bereiche möglich ist. Diese sind ebenfalls im generischen Modell zu berücksichtigen. Auf Grund der Vielfältigkeit der Platzierungsbedingung ist von einem kontinuierlichen Spektrum an Möglichkeiten auszugehen. Dies wird unter anderem durch die Vorgabe einer Infrastruktur weiter beeinflusst.

Bei der knotenbasierten Platzierung werden spezifische Möglichkeiten für die Standorte von Zentren vorgegeben. Die Platzierung eines Zentrums hat somit an einer vordefinierten Position zu erfolgen. Dazu wird in dem Modell eine entsprechende Liste dieser Positionen vorgehalten. Beim klassischen k-Center Problem entsprechen die Möglichkeiten den Standorten der Nutzer [65, 245]. In Spezialfällen wird davon abgewichen. Demgegenüber steht die Variante der freien Platzierung, wobei keine Einschränkungen in der Standortwahl vorgegeben sind. In diesem Fall ist die Liste der vordefinierten Möglichkeiten leer.

Die Mischformen werden über eine separate Schnittstelle im Modell handhabbar gemacht. Da hierbei kontinuierliche Bereiche auftreten, wird von einer freien Platzierung ausgegangen. Die eigentliche Bedingung bei einer Mischform wird über eine weitere Nebenbedingung abgebildet und schränkt damit die Platzierungsmöglichkeiten ein. Über dieses Prinzip der gestaffelten Nebenbedingungen lassen sich auch weitere Eigenschaften im generischen Modell ergänzen und abbilden. Ein mögliches Beispiel ist die Bevorzugung von bestimmten Regionen auf Grund von geringen Kosten, lokalen Subventionen oder ansässigen Fachkräften.

Hierzu werden einzelne Nebenbedingungen als Schichten aufgefasst, welche die Platzierung beeinflussen und zu beachten sind. Der Einfluss der Schichten erfolgt bei der Bewertung entsprechend einer Nutzwertanalyse additiv, damit der Effekt jeder Nebenbedingung sich entsprechend auswirkt. Jede Schicht besitzt auf Grund der Adressierung von verschiedenen Aspekten und Eigenschaften einen anderen Wertebereich y_x. Daher erhält jede Nebenbedingung einen Faktor f_x mit dem sich die Schicht auf das Gesamtsystem auswirkt. Somit ist es möglich den Einfluss der Schichten untereinander zu steuern und gegebenenfalls anzugleichen. Für eine Übersicht möglicher Einflussfaktoren auf die Standortwahl wird auf die wirtschaftswissenschaftliche Arbeit [51] verwiesen. Die Abbildung 5.8 verdeutlicht die Vorgehensweise bei der Modellierung von Nebenbedingungen und das Einbeziehen in die Standortwahl. Ein Standort von einem Zentrum lässt sich durch die Berechnung über die Schichten entsprechend bewerten. Ein Beispiel ist dazu in der Tabelle 5.3 dargestellt, wobei mehrere mögliche Standorte bei verschiedenen Nebenbedingungen miteinander verglichen werden. Die einzelnen Nebenbedingungen sind über eine Liste im Modell referenzierbar. Die hinter der Schnittstelle liegende Nebenbedingung ist mit den Eingangsdaten des jeweiligen Szenarios zu spezifizieren. Die Angabe der Bedingungen ist dabei flexibel in der Ausprägung. Hierbei lassen sich konkrete Werte vorgeben oder über eine

mehrdimensionale Funktion beschreiben, wobei der Wert von dem jeweiligen Standort abhängt. Auch Abhängigkeiten zu anderen Schichten sind über einheitliche Schnittstellen realisiert und somit in der Gesamtbewertung berücksichtigt.

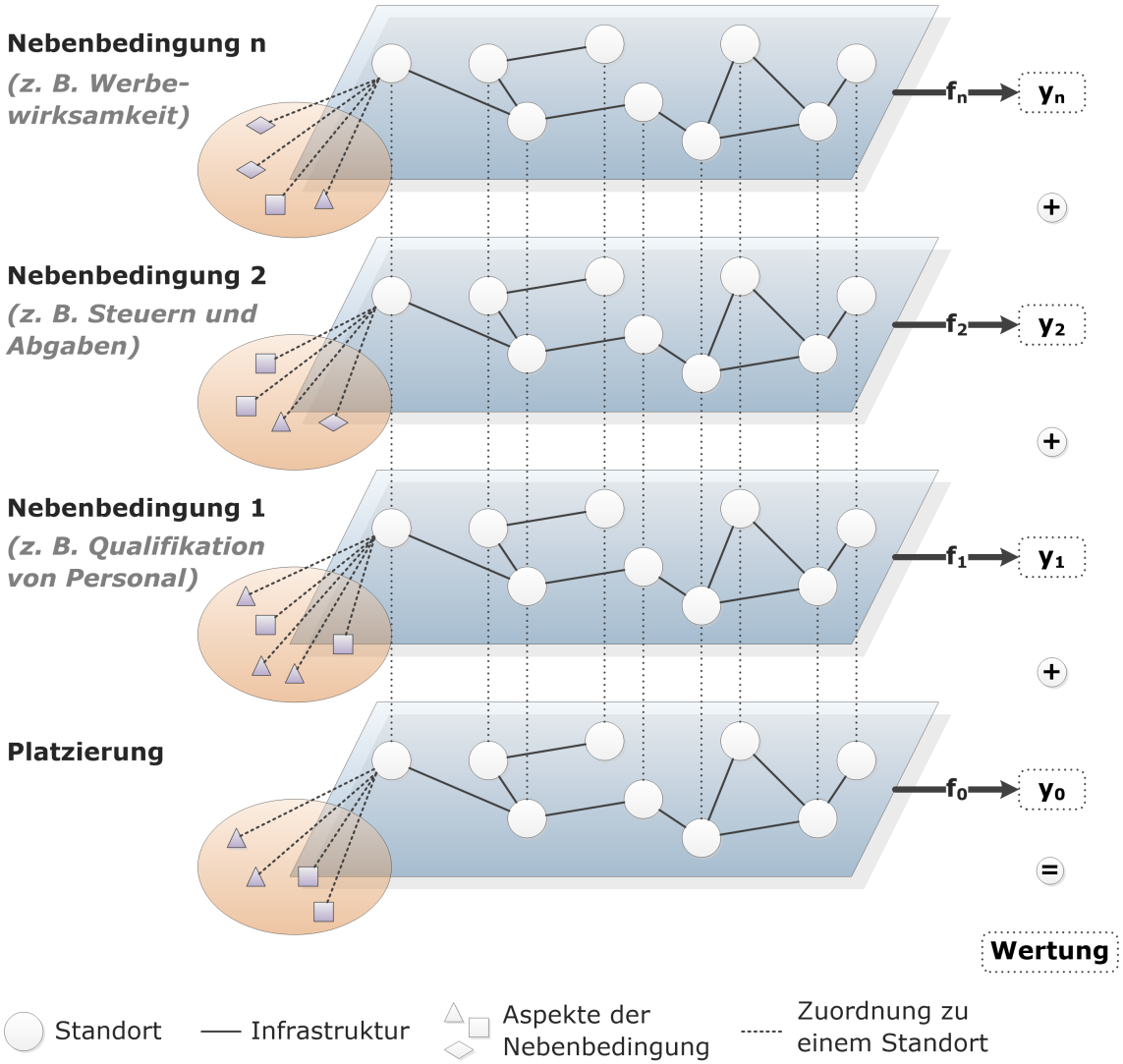


Abbildung 5.8: Auffassung der verschiedenen Nebenbedingungen als Schichten, welche sich additiv auf die Platzierung auswirken.

5.1.9 Zielfunktion

Gemäß der Problemstellung und dem Zusammenhang der Optimierungsaspekte in Abschnitt 2.3.5 gibt es mehrere Faktoren, welche die Qualität einer Lösung beeinflussen. Darüber hinaus sind in Abhängigkeit des Anwendungsgebietes unterschiedliche Zielfunktionen anzuwenden. Daher wird eine abstrakte Zielfunktion modelliert, die entsprechend zu spezifizieren ist und sich über Parameter steuern lässt. Unter Berücksichtigung der einzelnen Optimierungsaspekte Platzierung und Zuweisung ergeben sich dafür unterschiedliche Teilzielparameter.

Tabelle 5.3: Bewertung verschiedener Standorte anhand mehrerer Nebenbedingungen (Kleinere Werte sind besser.)

Nebenbedingung y_i	Faktor f_i	Standort 1	Standort 2	Standort 3
Infrastruktur	1	5,9	7,3	6,2
Arbeitsmarkt	0,5	2,1	1,5	1,7
Steuern	0,3	1,8	1,2	1,6
Staatliche Förderung	0,2	2	2	1,5
Politische Stabilität	1	1,1	1,3	2,5
Fläche zur Expansion	0,1	5,3	1,2	2,1
Gesamtwertung $\sum(f_i \cdot y_i)$	-	9,52	10,23	10,54

Platzierung

In Anbetracht eines gesamten Szenarios lässt sich für die Platzierung von Zentren und der einem Zentrum zugewiesenen Nutzer hinsichtlich verschiedener Kriterien optimieren. Dafür haben sich folgende Werte etabliert:

- MAX: Maximale Entfernung eines Nutzers zu seinem Zentrum
- AVG: Durchschnittliche Entfernung aller Nutzer zu deren jeweiligen Zentren
- Qx: Maximale Entfernung von einem definierten prozentualen Anteil X aller Nutzer gemäß dem x-ten Quantil zu deren Zentren

Neben der Betrachtung der Entfernungen zwischen den Nutzern und den Zentren sind in Abhängigkeit des Szenarios auch die Entfernungen zwischen den Zentren von Relevanz. Insbesondere bei Systemen gemäß dem Nabe und Speiche Prinzip ist dies zu berücksichtigen. Dafür bietet der generische Ansatz über einen Parameter die Möglichkeit für die Evaluation eines Modells diese Verbindungseigenschaft mit einzubeziehen. Durch die Konfiguration mittels eines Faktors lässt sich auch eine Wichtung dieses Aspektes vornehmen.

Zuweisung

Neben einer kurzen Entfernung sind die Zuweisungen der Anfragen und damit die Nutzerprofile für die Optimierung von Bedeutung. Bei ausschließlicher Betrachtung der Nutzerzuweisung unabhängig der Entfernungen besteht das Ziel die Nutzer primär den Zentren zuzuordnen, welche vorwiegend die Dienste im Interessen-Profil eines Nutzers anbieten. Somit sind die Nutzer hinsichtlich zueinander passender Profile zu optimieren. Die Abbildungen 5.9 und 5.10 veranschaulichen den Sachverhalt, indem die Histogramme der Profile gegenübergestellt sind. Zur Bewertung von zueinander ähnlicher Nutzerprofile bieten sich Cluster-Algorithmen und Korrelationsfunktionen an. Ein Zentrum verwaltet sein Speicher entsprechend den eintreffenden Anfragen, um diese zu beantworten. Demzufolge ergibt sich das Profil eines Zentrums aus den Anfragen der Nutzer und somit

aus deren Profilen. Hierbei ist zu beachten, dass an einem Zentrum mehrere Anfragen von unterschiedlichen Nutzern eintreffen.

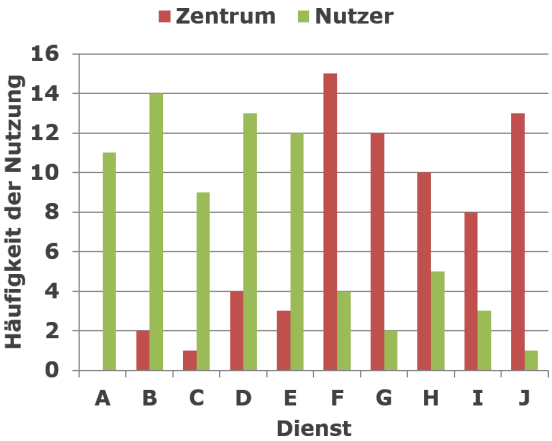


Abbildung 5.9: Negative Korrelation

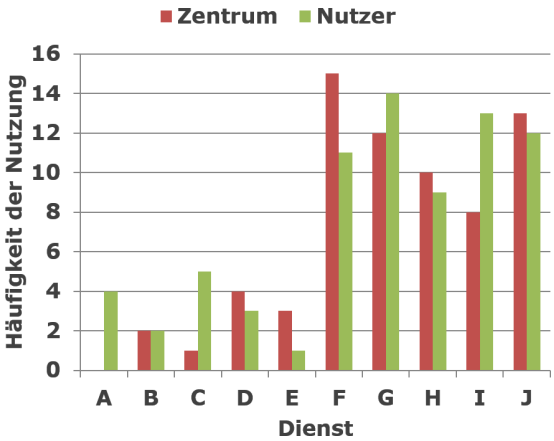


Abbildung 5.10: Positive Korrelation

Auf Grund der vielfältigen Nutzerinteressen ergibt sich ein breites Spektrum, wodurch sich ein komplexes Optimierungsproblem bezüglich der Gruppierung beziehungsweise Clusterung der Profile ergibt. Wenn jeder Dienst im System der Transportkette eine Dimension im multimodalen Suchraum erzeugt, wird der Abstand über eine Abstandsfunktion zwischen zwei Profilen auf Grund der hohen Dimensionalität bedeutungslos [246, 247]. Daher wird ein Ansatz über die Berechnung der Korrelationsfunktion gewählt, anstatt dafür direkt Cluster-Algorithmen einzusetzen.

Die Zuweisung eines Nutzer und damit dessen Profil ist dementsprechend hinsichtlich einer positiven Korrelation auszurichten, damit an einem Zentrum Anfragen nach Diensten eintreffen, welche auch im Speicher des Zentrums vorgehalten werden. Die Bewertung der Profilübereinstimmung erfolgt mittels Korrelationskoeffizient anhand der spezifizierten Zielfunktion der Gruppierung. Deren Wertebereich ist standardmäßig zwischen -1 und 1 normiert, wobei es sich bei -1 um inverse Datensätze handelt und bei 1 es identische Daten sind. Die Bewertung geschieht stets zwischen zwei Profilen. Hierfür ist vorab der Durchschnitt des an einem Zentrum aggregierten Profils zu bilden und zu normieren, um die Vergleichbarkeit zwischen einem Nutzerprofil und einem Zentrumprofil herzustellen. Eine örtliche Lokalität von ähnlichen Profilen um ein Zentrum begünstigt damit den Sachverhalt der Zuweisung und Standortoptimierung bezüglich kurzer Entfernungen.

Unter Nutzung der allgemeinen Zielfunktion wird eine abstrakte Korrelationsfunktion definiert. Die spezifische Ausprägung einer geeigneten Funktion steht in Abhängigkeit zum Anwendungsgebiet. Zu den gängigen Korrelationsfunktion gehören die folgenden:

- Kovarianz [248] mit $\mathrm{Cov}(X,Y) := \mathrm{E}\,W\big[(X-\mathrm{E}\,W(X))\cdot(Y-\mathrm{E}\,W(Y))\big]$
- Pearson [249] mit $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\sigma^2_{X,Y}}{\sigma_X \sigma_Y}$
- Kendall [250] mit $\tau = \frac{(\text{Anzahl gleicher paare})-(\text{Anzahl nicht gleicher paare})}{n(n-1)/2}$
- Spearman [251] mit $r_s = \frac{\sum_i (rang(x_i)-\overline{rang}_x)(rang(y_i)-\overline{rang}_y)}{\sqrt{\sum_i (rang(x_i)-\overline{rang}_x)^2}\sqrt{\sum_i (rang(y_i)-\overline{rang}_y)^2}} = \frac{\mathrm{Cov}(rang_x, rang_y)}{\sigma_{rang_x}\sigma_{rang_y}}$

Bedeutung der Variablen:

X,Y	...	Vektoren der Korrelation
n	...	Länge der Verkoren X, Y
EW	...	Erwartungswert
Var	...	Varianz mit $\mathrm{Var}(X) := \mathrm{E}\left((X-\mu)^2\right)$
rang(x)	...	Rang von x
$\overline{rang}_x$	...	Mittelwert der Ränge von x
σ_{rang_x}	...	Standardabweichung der Ränge von x mit $\sigma_X := \sqrt{\mathrm{Var}(X)}$

Die Modellierung von Interessen-Profilen der Nutzer erfolgt mittels Rängen gemäß Abschnitt 5.1.4. In Anbetracht dessen und der nachweislichen Korrelation mit der rangbasierten Zipf-Verteilung bietet sich die Verwendung von entsprechenden Korrelationsfunktionen zur Bewertung zusammengehöriger Profile an. Zu diesen gehören Pearson, Kendall und Spearman. Insbesondere bei der Korrelation nach Spearman kommt die Beziehung zwischen den verschiedenen Rängen zum Tragen, da diese Korrelationsfunktion eine vollständige Interrater-Reliabilität auf Ordinalskalenniveau vornimmt.

Das folgende Beispiel in Tabelle 5.4 erläutert die Berechnung des Pearson Korrelationskoeffizienten ρ. Beide Nutzer sind demselben Zentrum zugeordnet, wobei dieses das entsprechende aggregierte Profil besitzt.

Tabelle 5.4: Beispiel zur Rangkorrelation gemäß Pearson

Dienst	Nutzer 1	Nutzer 2	Zentrum
A	50 %	30 %	40 %
B	50 %	0 %	25 %
C	0 %	70 %	35 %
ρ	0,125	0,5	-

Speicherung

Neben den zuvor genannten statisch zu berechnenden Kriterien gibt es Werte, die sich nur über eine Simulation des Systems dynamisch ermitteln lassen. Diese sind emergent auf Grundlage der statischen Optimierung und stellen ein unabhängiges Zielkriterium dar. Dazu gehört die Auswirkung der Nutzeranfragen auf das Speicherverhalten in den Zentren einer für ein Szenario ermittelten Lösung. Die in diesem Zusammenhang entstehenden Effekte stellen ein Problem übergreifendes Optimierungskriterium dar. Dieses dient der umfänglichen Bewertung einer Lösung für eine Problemstellung und lässt sich auf Grund der NP-Vollständigkeit der Teilprobleme nicht effizient direkt optimieren.

Pareto-Front

Entsprechend einem multikriteriellen Optimierungsproblem lässt sich auf Basis des generischen Modells und der Simulation ein ganzheitliches Optimierungskriterium mit einer übergreifenden Zielfunktion bestimmen. Dies erfolgt entweder durch ein gewichtetes

Zusammenrechnen der einzelnen Teilzielparameter oder als eigenständige Bewertung. Im Hinblick des Einsatzzwecks einer Analyse von den Problemstellungen und einer Entscheidungsunterstützung wird für einen Eingabesatz an Parametern die sogenannte Pareto-Front ermittelt. In Anbetracht mehrerer Teilzielkriterien entspricht ein Pareto-Optimaler Zustand einer Lösung für ein Problem, bei der es nicht möglich ist ein Zielkriterium zu verbessern ohne gleichzeitig das andere zu verschlechtern. Die Menge der Pareto-Optimalen Zustände wird als Pareto-Front bezeichnet. Unter Berücksichtigung der NP-vollständigen Problemstellung und dem Einsatz von heuristischen Verfahren ergibt sich eine lokale Pareto-Front.

5.1.10 Modell-Erzeugung

Auf Grundlage der Eingangsdaten eines Szenarios entsteht über den generischen Ansatz das Modell der abstrakten Transportkette. Dazu werden die Daten entsprechend der im Kapitel 1.2.6 definierten Begrifflichkeiten kategorisiert und aufbereitet. Weiterhin erfolgt eine Sicherstellung der Voraussetzungen und Annahmen zur Anwendung des Modellierungsansatzes gemäß Kapitel 2.4. Die einzelnen Bestandteile des betrachteten Szenarios werden in die zuvor beschriebenen Komponenten modellgetrieben umgesetzt und über die spezifische Architektur zusammengeführt. Die maßgeblichen Bestandteile sind die einzelnen Nutzer mit deren Standort sowie dem Interessen-Profil. Darüber hinaus gehört die zu berücksichtigende Infrastruktur dazu.

Weiterhin werden im Modell hinsichtlich dem Ziel die zu analysierenden Werte gemäß Kapitel 2 festgelegt und die zu optimierenden Parameter für die Problemstellung definiert. Bezüglich der Lösungsberechnung erfolgt die Festlegung der Entfernungs- und Korrelationsfunktionen sowie der Zielkritierien. Dies schließt die Nebenbedingungen mit der Anzahl der geplanten Zentren und den Möglichkeiten zur Platzierung ein, wobei der spezifizierte Optimierungsansatz darauf zu achten hat. Zudem sind die Eigenschaften der durch das Transportsystem transferierten Dienste zu konfigurieren sowie die Parameter zur Speicherung und Verwaltung durch die Zentren festzulegen.

5.1.11 Optimierung

Basierend auf dem generischen Modell erfolgt die Lösungsberechnung und Optimierung. Bevor auf die verschiedenen Möglichkeiten der Optimierung eingegangen wird, erfolgt eine nähere Betrachtung des Optimierungsproblems und der Zielkriterien für eine optimierte Lösung. Allgemein gibt es in Anbetracht der Problemstellung gemäß Kapitel 1.4 zwei Zielkriterien. Aus Sicht des Nutzers besteht ein Interesse in der Minimierung der Antwortzeit bei hoher QoS. Dies wird durch kurze Strecken zwischen Nutzer und Zentrum erreicht sowie einer sofortigen Bereitstellung des angeforderten Dienstes durch das Zentrum. Aus Sicht des Betreibers ist der Einsatz der Ressourcen zu minimieren und die Systemauslastung bei hoher Bereitstellung einer QoS zu verringern. Dafür hat eine effiziente Nutzung der Zwischenspeicher bei den Zentren zu erfolgen und es müssen kurze Strecken bei der Bereitstellung und Weitergabe der Dienste gefunden werden. Somit besteht das Optimierungsproblem grundlegend aus der Verringerung

der Entfernung hinsichtlich einer reduzierten Transferzeit und Last auf die Infrastruktur. Darüber hinaus ist eine angepasste Zuweisung der Anfragen nach gleichen Diensten an einem Zentrum anzustreben, um eine optimierte Abarbeitung und Speicherung zu ermöglichen, was wiederum eine spezifische Anpassung und leichtere Automatisierung von Prozessschritten zulässt.

Zur Verdeutlichung des Optimierungsproblems der Anfragezuweisung ist in der Abbildung 5.11 ein Beispiel gegeben. Nutzer 1 und 2 fragen stets den orangenen Dienst (A) an und sind dem Zentrum 1 mit einem Speicherplatz für einen Dienst zugewiesen. Nutzer 3 fragt zur gleicher Wahrscheinlichkeit den orangenen oder den grünen Dienst (A, B) an und ist dem Zentrum 2 mit ebenfalls einem Speicherplatz zugewiesen. Der Nutzer 4 besitzt zum Zentrum 1 eine kürzere Entfernung als zum Zentrum 2 und verlangt stets nach dem grünen Dienst (B). Für die regelmäßige Zuweisung der Anfragen ist die Entscheidung zu treffen, an welches Zentrum die Anfragen des Nutzers 4 zu senden sind, um für den Nutzer und das Gesamtsystem eine optimierte Situation zu erreichen. Eine Anfrage nach einem Dienst am Zentrum, welcher im Speicher vorgehalten wird, lässt sich entsprechend besser beantworten. Das Ziel besteht demzufolge in der Ausnutzung der Zwischenspeicher zur Erreichung der Domänenvorteile. Mit einer passenden Zuweisung der Anfrage eines Nutzers an ein Zentrum nach einem Dienst, welcher bereits zuvor angefragt wurde, lässt sich das Gesamtsystemverhalten optimieren.

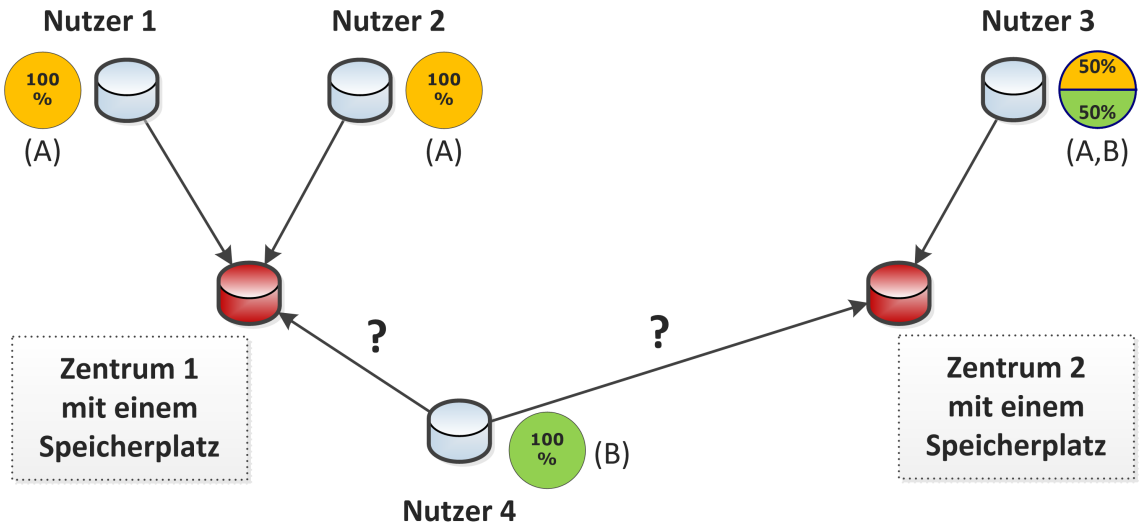


Abbildung 5.11: Beispiel für die Problemstellung einer optimierten Anfragezuweisung

Auf Grund der verschiedenen Zielkriterien der Problemstellung und der unterschiedlichen Interessen von den Dienstleistern bestehen mehrere Ziele für eine optimierte Lösung. Bei einem solchen multikriteriellen Optimierungsproblem besteht das Ziel nicht ausschließlich darin eine optimierte Lösung für ein Zielkriterium zu berechnen, sondern Lösungen zu erhalten, die mehrere Kriterien gleichzeitig möglichst optimal erfüllen. Demzufolge ist das Ziel der Optimierung mehrere Lösungen zu ermitteln, die hinsichtlich verschiedener Zielkriterien ein optimiertes Ergebnis liefern. Diese pareto-optimalen Ergebnisse bilden die sogenannte Pareto-Front.

Gemäß dem Zusammenhang der Optimierungsaspekte in Kapitel 2.3.5 erfolgt hier eine Betrachtung der vier Parameter: Anzahl der Zentren, Platzierung der Zentren, Zuweisung der Nutzer und Speicherung der Dienste. Zur Reduktion der Komplexität erfolgt die Identifizierung der notwendigen Anzahl von Zentren über eine mehrfache Ausführung der Optimierung mit unterschiedlichen Vorgaben für die Anzahl der Zentren. Weiterhin wird die Speicherung der Dienste sowie der damit im Zusammenhang stehenden Optimierungsaspekte über die Simulation ermittelt. Daher lässt sich eine qualitative Aussage über eine Lösung erst nach der Entscheidung über die Platzierung der Zentren und Zuweisung der Nutzer berechnen. Der zentrale Ansatz des Optimierungsproblems besteht somit in der Berechnung von optimierten Lösungen hinsichtlich der Platzierung der Zentren und der entsprechenden Zuweisung der Nutzer. Die Pareto-Front besteht damit aus zwei Zielkriterien. Zum einen ist eine Minimierung der Entfernung von einem Nutzer zu seinem Zentrum zu erreichen. Zum anderen besteht das Ziel in der Maximierung hinsichtlich einer positiven Profil-Korrelation der an einem Zentrum zugewiesenen Nutzer.

In diesem Zusammenhang lässt sich die Optimierung hinsichtlich statischer und dynamischer Kriterien einteilen. Bei der statischen Optimierung wird auf Grundlage der Eingangsdaten eine initiale Lösung ermittelt, welche keine stochastischen und zufälligen Einflüsse berücksichtigt. In der dynamischen Optimierung haben die Ergebnisse der Simulationen Einfluss auf die Lösungsermittlung. Das multikriterielle Optimierungsproblem lässt sich auf zwei verschiedene Weisen realisieren. Der in der vorliegenden Arbeit entwickelte Ansatz sieht auf Grund der NP-vollständigen Teilproblemstellungen eine sequentielle Optimierung hinsichtlich der verschiedenen Teilzielkriterien vor. Darüber hinaus besteht die Möglichkeit mit bestehenden Optimierungsansätzen direkt eine optimierte Lösung für die verschiedenen Zielkriterien zu errechnen.

Sequentielle Optimierung

Bei der sequentiellen Optimierung erfolgt eine schrittweise Verbesserung einer Lösung hinsichtlich der verschiedenen Zielkriterien. Ausgehend von einer initialen Lösung wird diese abgewandelt und verändert, um neue Lösungen zu erhalten. Die verschiedenen Zielkriterien werden dabei sequentiell betrachtet und dahingehend zielgerichtet optimiert, sodass jeweils eine Verbesserung bezüglich eines gewählten Zielkriteriums erreicht wird. Beginnend beim zentralen Optimierungsproblem für eine Pareto-Front hat eine Optimierung hinsichtlich der optimalen Platzierung der Zentren oder der optimalen Zuweisung für eine Gruppeneinteilung zu erfolgen. In Anbetracht der Problemstellung für eine optimierte Transportkette ist die Infrastruktur die bestimmende Eigenschaft zum Aufbau des Transportsystems. Dies wirkt sich vorwiegend auf die Platzierung der Zentren und die Entfernungen in Relation zu den Nutzern aus. Daher erfolgt zur initalen Lösungsfindung eine Platzierung der Zentren anhand der optimierten Entfernungen, wobei die Nutzerinteressen keine Beachtung finden. Dadurch wird ein optimiertes Ergebnis bezüglich der Platzierung erreicht, welches somit das eine Ende der Pareto-Front darstellt. Dies entspricht auch dem Fall, dass alle Nutzer die gleichen Interessen aufweisen bzw. eine hohe Lokalität um das jeweilige Zentrum besteht. Zur Erreichung der initialen Lösung lassen sich die bestehenden Verfahren gemäß Kapitel 4.8 verwenden. Insbesondere ist dafür der in dieser Arbeit entwickelte Optimierungsansatz vorgesehen.

Von dieser initialen Lösung ausgehend, wird anschließend in einem iterativen Prozess nach Lösungen hinsichtlich eines der anderen Zielkriterien gesucht. Hierbei werden die Nutzerprofile mit in die Optimierung einbezogen, indem über die spezifizierte Korrelationsfunktion eine Bewertung zueinander passender Profile erfolgt. Dazu wird die Lösung schrittweise für eine optimierte Profil-Korrelation abgeändert und die Zwischenlösungen mit aufgezeichnet. Der Wert der Korrelation lässt Rückschlüsse auf das zu erwartende Ergebnis zu, sodass der Optimierungsalgorithmus eine Entscheidung treffen kann. Zur Abänderung einer bestehenden Lösung bieten sich heuristische Optimierungsverfahren wie Greedy oder Random Walk an. Weiterhin lassen sich auch die Cluster-Verfahren gemäß Kapitel 4.8 für diesen Ansatz adaptieren. Insbesondere durch die einheitliche Grundlage des generischen Modells lässt sich dieses sequentiell durch verschiedene Algorithmen optimieren.

Der in der vorliegenden Arbeit entwickelte Ansatz verwendet einen angepassten Greedy-Algorithmus, um die anfängliche Lösung hin zu einer Profilkorrelation zu optimieren. Dazu erfolgt eine Verbesserung der Zuweisung von Nutzern zu deren bevorzugten Zentrum. Daher fokussiert der Algorithmus auf die Erhöhung des Korrelationskoeffizienten bei jedem Zentrum. Der Standort der Zentren richtet dabei an den entsprechend zugewiesenen Nutzern. Eine Neuzuordnung eines Nutzers hat einen Einfluss auf die Korrelation am bisherigen Zentrum und dem neuen Zentrum. Der Koeffizient repräsentiert dabei die Gleichheit aller Nutzeranfragen, die an das selbe Zentrum gerichtet sind. Daher wird das Profil des betrachteten Nutzers zum potentiellen Profil des Zentrums bei einer entsprechenden Neuzuweisung mit integriert. Für die Bewertung der Korrelation wird eine der im Abschitt 5.1.9 beschriebenen Funktionen verwendet. Insbesondere bietet sich für den betrachteten Sachverhalt der Rangkoeffizient von Spearman an.

In jeder Iteration berechnet der Algorithmus für jeden Nutzer den Korrelationskoeffizienten mit jedem Zentrum. Wenn der Korrelationskoeffizient für das neue Zentrum positiv ist, vergleicht dieser den Wert mit dem Koeffizienten mit dem aktuellen Zentrum. Die Nutzer mit einem verbesserten Korrelationskoeffizienten werden am Ende jeder Iteration neu zugeordnet. Dabei wird die veränderte Entfernung eines Nutzers mit in Betracht gezogen. Die Zuweisung aller Nutzer erfolgt dabei gleichzeitig zum neuen Zentrum mit dem besten Kompromiss aus Korrelationskoeffizienten und Entfernungen. Dieser Vorgang wird solange wiederholt, bis keine Neuzuordnung stattfindet. Da das Verfahren nur bessere Konstellationen akzeptiert, terminiert der Optimierungsprozess stets.

Gleichzeitige Optimierung mittels multikriteriellen Optimierungsalgorithmen

Im Hinblick auf eine gleichzeitige Mehrzieloptimierung sind mehrere Zielsetzungen miteinander zu vereinen. Dabei lassen sich voneinander, grundsätzlich unabhängige Zielkriterien nicht immer zu einem gemeinsamen Ziel zusammenfassen. Sofern sich die verschiedenen Zielkriterien zusammenfassen lassen oder eine Hierarchie existiert, dann besteht die Möglichkeit ein übergeordnetes Zielkriterium zu definieren. Dazu sind die verschiedenen Kriterien mittels Gewichtsfaktoren über eine gemeinsame Funktion abgebildet. Dadurch wird das Problem vereinfacht, sodass nur noch ein Ziel besteht und damit die multikriterielle Optimierung über die Verbesserung des Gesamtkriteriums

erreicht wird. Anderenfalls müssen die Kriterien separat betrachtet werden. Etablierte Optimierungsverfahren für solche multikriteriellen Problemstellungen sind *Archived Multi Objective Simulated Annealing (AMOSA) [252]* und *Non-dominated Sorting Genetic Algorithms (NSGA) [217]*. Hierbei werden ausgehend von einer Menge zufälliger, initaler Lösungen über Mutationen neue Lösungen generiert. Die gleichzeitige oder abwechselnde Mutation hinsichtlich geänderter Platzierung der Zentren und modifizierten Zuweisung der Nutzer zu Zentren ermöglicht eine multikriterielle Optimierung.

5.1.12 Simulation

Zur Analyse von Prozessen und dynamischen Zusammenhängen erfolgt eine Simulation der Transportkette. Dies ist insbesondere bei stochastischen Einflüssen wie den Anfragen auf Grundlage der Interessen-Profilen von den Nutzern von Bedeutung. Mit der Simulation lassen sich auch die berechneten Lösungen zu statischen Optimierungskriterien auswerten. Entsprechend des generischen Ansatzes muss die Simulation dazu die verschiedenen Verhaltensweisen einer Transportkette in den unterschiedlichen Anwendungsgebieten nachbilden. Daher erfolgt die Simulation auf Grundlage des generischen Ansatzes mit aus dem Szenario erstellten Modell. In den beiden betrachteten Anwendungsgebieten sind zwei grundsätzliche Arbeitsweisen der Zentren zu unterscheiden. Diese sind das iterative und rekursive Systemverhalten der Zentren, welches auch Einfluss auf das Gesamtsystem ausübt. Die unterschiedlichen Verhaltensweisen sind in der Simulation durch funktionale Eigenschaften der Zentren abzubilden. Ein Zentrum beantwortet hierbei die Anfragen der Nutzer unverzüglich. Zur weiteren Präzisierung der Simulation ist das Verhalten über ein Warteschlangenmodell zu realisieren. Mittels dem $M/M/c/k$ -Modell [253] lassen sich Verarbeitungszeiten in einem Zentrum abbilden und mit in das Systemverhalten einbeziehen.

Iteratives Systemverhalten

Bei einem iterativen bzw. nicht kooperativem Systemverhalten sendet ein Nutzer seine Anfrage an ein Zentrum und erhält von diesem eine positive oder negative Antwort. Wenn die Antwort positiv ausfällt, dann enthält diese den angeforderten Dienst und der Nutzer muss kein weiteres Zentrum anfragen. Bei einem negativen Bescheid hat das angefragte Zentrum den angeforderten Dienst momentan nicht im Zwischenspeicher und kann diesen nicht zur Verfügung stellen. Daraufhin muss der Nutzer selbst weitere Zentren anfragen, bis dieser den ausgewählten Dienst erhält. Dies erfolgt sequentiell. Falls alle Zentren den Dienst nicht zur Verfügung stellen, dann wird die Quelle um eine Antwort aufgefordert.

Rekursives Systemverhalten

Beim rekursiven Systemverhalten unterstützt das System den Nutzer bei der Erlangung der gewünschten Antwort. Es wird daher auch kooperatives Verhalten genannt. Dabei übernimmt das angefragte Zentrum die Rolle einer Quelle und sorgt für die Bereitstellung der Dienste. Die Anfrage eines Nutzers wird somit stets positiv mit dem geforderten Dienst

beantwortet. Dazu schaut das Zentrum zunächst im eigenen Speicher nach und sendet bei Verfügbarkeit den Dienst sofort an den Nutzer. Wenn das Zentrum den Dienst selbst nicht im Zwischenspeicher zur Verfügung hat, dann fragt es weitere Zentren bis hin zur Quelle an. Anschließend wird der angeforderte Dienst beim Zentrum im Zwischenspeicher abgelegt und anschließend dem anfragenden Nutzer zugesendet. Hierbei ist die maximale Kapazität des Speichers zu beachten.

Nabe und Speiche Systemverhalten

Unter Nutzung der vorausgegangenen Beschreibung des rekursiven Systemverhaltens wird die Simulation von Nabe und Speiche Systemen ermöglicht. Dabei übernehmen die Zentren die Rolle der Hubs, über welche die Nutzer deren Dienst an andere Nutzer im System transferieren. Ein Nutzer stellt hierbei somit gleichzeitig eine Quelle für einen Dienst dar. Den Zentren ist bei diesem Systemverhalten bekannt, welcher Nutzer über welches Zentrum zu erreichen ist. Hierbei besteht die Verwaltung des Zwischenspeichers eines Zentrums darin, die eintreffenden Anfragen in einer Warteschlange aufzunehmen und zu sortieren. Anschließend sind die Informationen an das zugehörige Zentrum weiterzuleiten. Nach dem selben Prinzip werden auch die Antworten verarbeitet. Somit erfolgt der Transfer von Anfragen und Diensten für das Zentrum transparent. Die Nutzer fragen bei diesem Verhalten vorrangig nach Diensten, welche im System bekannt sind.

Verwaltung des Zwischenspeichers

Bei der Verwaltung des Speichers sind die Eigenschaften der in einer Domäne betrachteten Dienste zu berücksichtigen. In Abhängigkeit von diesen ist die Verwaltungsstrategie des Zwischenspeichers zu wählen. Hierbei sind zwei grundsätzliche Arten zu unterscheiden. In Anlehnung an die Logistik sind bei physischen Diensten vorausschauende Bestell- und Bestandsoptimierungsstrategien anzuwenden. Im Bereich von Rechnernetzen erfolgt bei virtuellen Diensten die Nutzung von Ersetzungsstrategien.

Bestandsoptimierungsstrategien - Physische Dienste: Bei der Verwendung von Bestandsoptimierungsstrategien werden die Dienste zeitlich vor einer Anfrage im Zentrum bereit gehalten, da ansonsten davon ausgegangen wird, dass eine zeitnahe Beantwortung nicht möglich ist. Ein vom Zentrum ausgelieferter Dienst reduziert den Vorrat und somit die verfügbare Anzahl im Zentrum für weitere Anfragen. Mit jeder Auslieferung wird daher Speicherkapazität entsprechend den Diensteigenschaften frei gegeben. Das Ziel der Bestandsoptimierungsstrategien besteht darin, soweit möglich sicher zu stellen, dass ein spezifischer Dienst im Zentrum zur Verfügung steht, wenn eine Anfrage dafür eintrifft. Die vorgehaltene Anzahl ist anhand des Bedarfs auszurichten, um die notwendige Speicherkapazität gering zu halten. Dazu erfolgt in Abhängigkeit der freien Kapazität eine Bestellung der Dienste bei den Quellen, welche Dienste aus Sicht des Zentrums demnächst in Art und Anzahl benötigt werden. Ein einmal im Zentrum gespeicherter Dienst ist durch die Anfragen der Nutzer zu verbrauchen und kann nicht bei verbleibender Restmenge durch einen anderen Dienst ersetzt werden.

Die Bestellpolitik regelt zu welchem Zeitpunkt eine Bestellung aufgegeben wird und welche Mengen angefordert werden. Hierbei sind die Fixkosten wie die Verteilungs- und Transportkosten sowie die Lagerhaltungskosten gegeneinander abzuwägen. In der Bestellpolitik werden die folgenden zwei grundlegenden Verfahren unterschieden [254–258]:

- Bestellpunktsystem (Mengenorientierung): Auslösen einer Bestellung bei Unterschreitung einer Vorratsgrenze
- Bestellrhythmussystem (Terminorientierung): Regelmäßige Bestellung in definierten Zeitabständen zum Auffüllen des Bestandes bis zur Sollkapazität

Darüber hinaus gibt es verschiedene Mischformen von Bestandsoptimierungsstrategien. Eine ausführliche Vorgehensweise ist in [259] beschrieben, wobei viele verschiedene Bedingungen beachtet werden. Unter anderem sind die Vorgaben der Quellen mit in die Planung einzubeziehen und die Strategie entsprechend daran anzupassen. Die optimale Bestellstrategie ist abhängig von dem jeweilig einzukaufenden Dienst. Neben der Bestandsoptimierung ist entsprechend der verfügbaren Kapazität über die angebotenen Dienste zu entscheiden. Hierbei gibt es in jedem Zentrum ein vorab definiertes Angebot von Diensten, welches sich in Abhängigkeit des Bedarfs auch ändern lässt. Falls eine Anfrage nach einem Dienst an einem Zentrum eintrifft und das Zentrum diesen nicht zur Verfügung hat, dann ist diese Anfrage an ein anderes Zentrum zu stellen bzw. zu einem späteren Zeitpunkt zu erfüllen.

Ersetzungsstrategien - Virtuelle Dienste: Beim Einsatz von Ersetzungsstrategien wird davon ausgegangen, dass ein im Zentrum gespeicherter Dienst sich ohne weitere Folgen entfernen lässt und somit den Speicherplatz freigibt [260]. Insbesondere beim Erreichen der maximalen Speicherkapazität besteht damit die Flexibilität einen vorrätigen Dienst durch einen neuen zu ersetzen, welcher mehr benötigt wird. Allgemein besteht das Ziel der Ersetzung darin, möglichst genau die Dienste bei einem Zentrum vorzuhalten, welche auch zukünftig angefragt werden, um diese sofort zu bedienen. Zur Verwaltung und zur Auswahl des Speichers gibt es verschiedene Strategien, wobei die bedeutendsten im Folgenden aufgelistet sind:

- **Béládi - Optimum** [261]
  Unter Berücksichtigung aller zukünftigen Anfragen wird das Element aus dem Speicher durch das Neue ersetzt, welches am längsten nicht benötigt wird.
- **Sequentiell**
  Die Speicherplätze sind organisatorisch entsprechend einem Kreis angeordnet und werden über einen Zeiger nacheinander vergeben und ersetzt.
- **Zufall**
  Der Speicherplatz für ein neues Element wird zufällig ausgewählt.

- **First In First Out (FIFO)** [262]
  Wenn die Speicherkapazität ausgeschöpft ist und ein weiteres Element vorzuhalten ist, dann wird stets das zuerst eingefügte Element ersetzt unabhängig der Anzahl der Zugriffe.
- **Clock** [263]
  Die Speicherplätze sind organisatorisch entsprechend einem Kreis angeordnet und werden im Uhrzeigersinn vergeben, wobei der Zeiger auf das zuletzt genutzte Element zeigt.
- **Least Recently Used (LRU)** [264]
  Die Verhaltensweise entfernt das Element, welche am längsten nicht angefordert wurde.
- **Least Recently Used k (LRU-k)** [265]
  Die Handlungsweise betrachtet im Gegensatz zu LRU nicht den letzten Zeitpunkt, sondern speichert für jedes Element eine Historie und nutzt zur Entscheidung den i-ten Zeitstempel in der Historie von jedem Element.
- **Least Recently Used with Probability q (q-LRU)** [39]
  Dieses Verfahren speichert ein neues Element nur mit der Wahrscheinlichkeit q und verhält sich entsprechend dem LRU Verfahren.
- **Least Frequently Used (LFU)** [264]
  Die Regel hält die Elemente vor, auf welche statistisch am häufigsten zugegriffen wird.
- **Least Recently/Frequently Used (LRFU)** [266]
  Dieser hybride Ansatz kombiniert LRU und LFU miteinander, wobei automatisch zwischen den Algorithmen balanciert wird.
- **Frequency-Based Replacement (FBR)** [267]
  Der Algorithmus kombiniert Zugriffshäufigkeit und Alter zur Entscheidung über das zu verdrängende Element aus dem Speicher. Dieser wird dazu in drei Abschnitte unterteilt. Die Elemente wandern in Abhängigkeit der Zugriffshäufigkeit durch die Altersabschnitte, bis diese im letzten Abschnitt durch ein neues Element im ersten Abschnitt ersetzt werden.
- **Multi Queue (MQ)** [268]
  Der Speicher wird in mehrere Warteschlangen eingeteilt, welche nach dem LRU Prinzip funktionieren. Entsprechend der Zugriffshäufigkeit werden die Elemente zwischen den Warteschlangen verschoben.
- **Adaptive Replacement Cache (ARC)** [269] Der hybride Ansatz balanciert zwischen LRU und LFU mit einer automatischen Anpassung in Abhängigkeit des gegenwärtigen Anfrageverhaltens.
- **Clock with Adaptive Replacement (CAR)** [270]
  Der Ansatz kombiniert ARC und Clock, wobei eine automatische und kontinuierliche Anpassung der Parameter stattfindet.

- **Low Inter-reference Recency Set (LIRS)** [271]
  Unter Nutzung des Referenzierungsabstandes der einzelnen Elemente erfolgt die dynamische Bestimmung des zu verdrängenden Speicherplatzes.
- **Most Recently Used (MRU)** [264]
  Es wird das Element auf dem Speicher verdrängt, welches bisher am häufigsten genutzt wurde.

Zudem gibt es eine Master-Policy [272], welche mehrere Strategien simuliert und in Abhängigkeit der Leistung dynamisch wechselt.

Unter der Annahme, dass alle Informationen über zukünftige Anfragen an ein Zentrum vorliegen, lässt sich gemäß der Strategie von Bélády der zur Verfügung stehende Speicherplatz optimal verwalten. Dabei werden zu einem Zeitpunkt die Dienste im Speicher verdrängt, welche in Zukunft am längsten nicht benötigt werden. Da zukünftige Anfragen jedoch oft in der Realität nicht bekannt sind, versuchen die anderen Strategien trotz des unvollständigen Wissens möglichst nah an das optimale Verhalten heran zu kommen. Bei einer gleichmäßigen Verteilung der angefragten Dienste an einem Zentrum ergibt sich für die meisten Ersetzungsstrategien eine ähnliche Performance bzw. schlechter als bei LRU [273]. Da jedoch die Anfrageverteilung von den jeweiligen Profilen abhängt, ist die passende Ersetzungsstrategie von dem jeweiligen Szenario und des Anwendungsbereiches abhängig.

Weiterhin ist zu beachten, dass eine Vergrößerung der Speicherkapazität nicht automatisch zu einer Steigerung der Trefferquote führt. Gemäß dem Phänomen von Bélády's Anomalie zeigt sich, dass bei nicht optimalen Strategien durch ein um eins vergrößerter Speicher eine annähernde Verdopplung der Menge an Speicherfehlern entstehen kann [141, 274].

Initialisierung

In Abhängigkeit des spezifizierten Systemverhaltens und der Reaktion von Zentren auf die Anfragen der Nutzer ist die Initialisierung der Dienste im Speicher der Zentren zu berücksichtigen. Zu Beginn der Simulation des Systems wird von einem leeren Speicher der Zentren ausgegangen. Zur Initialisierung bieten sich zwei verschiedene Möglichkeiten. Entweder werden im Zentrum manuell vorab Dienste unter Berücksichtigung der maximalen Speicherkapazität festgelegt, oder es erfolgt eine automatische Initialisierung während der Laufzeit. Bei einer manuellen Spezifizierung der vorgehaltenen Dienste wird für jedes Zentrum einzeln definiert, wie der Speicher belegt ist. Hierbei bietet es sich an, die Dienste an dem jeweiligen Zentrum vorzuhalten, welche am ehesten von der zugeordneten Nutzergruppe angefragt werden. Dazu sind die Profile der entsprechenden Nutzergruppe zu betrachten und die Dienste anhand deren aggregierten Wahrscheinlichkeiten aufzulisten. Die Dienste mit der höchsten Anfragewahrscheinlichkeit werden anschließend im Zentrum bis zur maximalen Speicherkapazität abgelegt. Bei einer automatischen Initialisierung während der Laufzeit ist der Speicher zu Beginn der Simulation nicht belegt und das Zentrum reagiert entsprechend der freien Kapazität des Speichers. In Abhängigkeit der Analyse sind die statistischen Werte in Verbindung

mit dem Speicher zurück zu setzen, um den Einfluss der Initialisierung aus zu lassen. In Bezug auf das Systemverhalten hat beim iterativen Systemverhalten die Art der Initialisierung einen direkten Einfluss auf die Systemleistung, da der Inhalt des Speichers als statisch angesehen wird. Beim rekursiven Systemverhalten genügt eine automatische Initialisierung des Speichers, da dieser dynamisch an die Nutzeranfragen angepasst wird.

Simulation von Nutzeranfragen

Zur Nachbildung des Systemverhaltens stellen die Nutzer Anfragen an die Zentren. Dabei bestehen hauptsächlich zwei Möglichkeiten, um aus dem Profil eines Nutzers Anfragen zu generieren. Ein Nutzer kann sich entsprechend einem vorgegebenen Trace verhalten. Hierbei wird die Historie wiederholt, indem ein Nutzer in der gleichen Abfolge und Zwischenabständen Anfragen an ein Zentrum stellt, wie zuvor aufgezeichnet. Bei Bedarf lässt sich der Startpunkt innerhalb des Traces variieren.

Demgegenüber besteht die Möglichkeit Anfragen gemäß dem Interessen-Profil und der wahrscheinlichkeitsbedingten Anfrageverteilung zu erstellen. Hierbei ist zu Beginn die Entscheidung zu treffen, ob ein betrachteter Nutzer eine Anfrage stellt oder nicht. Dazu wird der Parameter der Anfragehäufigkeit berücksichtigt und in die Zufallsentscheidung mit einbezogen. Falls der Nutzer keine Anfrage stellt, ist der Prozessschritt beendet. Bei einer Anfrage ist zwischen verschiedenen Kategorien zu unterscheiden, welche über Parameter der Simulation konfigurierbar sind. Dabei werden die folgenden Arten unterschieden:

- Nutzer handelt gemäß Profil
- Nutzer fragt nach Dienst in der Domäne ohne denen im Profil
- Nutzer stellt eine zufällige Anfrage nach einem bekannten Dienst
- Nutzer stellt eine Anfrage nach einem unbekannten Dienst

Die Selektion erfolgt gemäß der Parameter über eine beeinflusste Wahrscheinlichkeitsauswahl.

Nutzer handelt gemäß Profil: Wenn ein Nutzer sich entsprechend seinem Profil verhält, dann stellt dieser eine Anfrage nach einem Dienst innerhalb seines Interessenbereiches. Dazu wird die Tabelle der Anfrageverteilung verwendet. Zur Auswahl des Dienstes werden die einzelnen Wahrscheinlichkeiten im Profil berücksichtigt und über eine beeinflusste Zufallsentscheidung ausgewählt.

Nutzer fragt nach Dienst in der Domäne ohne denen im Profil: Bei dieser Kategorie von Anfragen wählt der Nutzer einen Dienst aus der Menge der in einer Domäne vorhandenen Dienste aus, wobei der Dienst nicht im Profil des Nutzers liegt. Somit wird die Differenzmenge aus den Diensten in der Domäne ohne den Diensten im Interessenbereich gebildet. Anschließend erfolgt die zufällige Entscheidung aus der restlichen Menge

über den anzufragenden Dienst. Somit lässt sich analysieren, inwieweit eine Lösung von den Interessen-Profilen abhängt und stabil gegenüber weiteren unvorhergesehenen Ereignissen ist.

Nutzer stellt eine zufällige Anfrage nach einem bekannten Dienst: Diese Art der Verhaltensmodellierung entspricht dem standardmäßigen Verhalten von Nutzern, wenn keine Informationen über die Interessens-Profile der Nutzer vorliegen. Hierbei wird aus der Menge an Diensten der Domäne gleich verteilt ein zufälliger ausgewählt. Dadurch lässt sich der Unterschied zu dem Verhalten gemäß Profil analysieren und die damit verbundenen Auswirkungen angeben.

Nutzer stellt eine Anfrage nach einem unbekannten Dienst: Um den Einfluss von unvorhergesehenen Anfragen auf das System simulieren zu können, ist im Modell eine weitere Verhaltensweise modelliert. Ein Nutzer hat dafür die Möglichkeit eine Anfrage nach einem unbekannten Dienst in der Domäne zu stellen. Damit lässt sich auch eine Veränderung der in der Domäne vorhandenen Dienste behandeln.

5.2 Spezifischer Anwendungsbereich für das Modell

Zur Verdeutlichung der flexiblen Nutzbarkeit des generischen Modellierungsansatzes ist in diesem Abschnitt der Einsatz im Bereich Logistik und Rechnernetze beschrieben. Dazu wird jeweils für ein spezifisches Szenario die schrittweise Adaption dargelegt. Darüber hinaus erfolgt die Nutzung des generischen Modells für wissenschaftliche Analysen im Rahmen der Geolokalisation. Dazu wird ein weiteres Modell entwickelt, welches spezifisch für diesen Anwendungszweck ist und Fähigkeiten des generischen Modells verwendet.

5.2.1 Spezifisches Modell für Logistik

Dieser Abschnitt beschreibt die flexible Anwendung des generischen Modells für ein Szenario der Logistik. Auf die spezifischen Eigenschaften des Anwendungsgebietes wird an der jeweiligen Stelle bei der Erläuterung des entsprechenden Aspektes eingegangen. Die hier betrachtete Transportkette besteht aus einem zwei-stufigen Prozess. Die zu transportierenden Waren einer Firma werden einem logistischen Netzwerk übergeben, welches den Transport vom Herstellungsort über die Zwischenlager an die lokalen Händler übernimmt. Hierzu erfolgt die Planung von den zentralen Zwischenlagern zur weiteren Verteilung an die Verkaufsorte.

Im Modell stellen die Verkaufsorte die Nutzer dar, welche durch die Zwischenlager zu beliefern sind. Die Geokoordinaten von diesen Standorten werden mit dem Szenario vorgegeben. Die durch die Transportkette zu transferierenden Waren sind im generischen Modell durch die Dienste abgebildet, welche in dedizierten Mengen an die Nutzer abgegeben werden. Der Transport erfolgt dabei über die Infrastruktur des Straßennetzes. Dieses stellt somit die Grundlage für die Entfernungsermittlung zwischen

den verschiedenen Standorten dar. In Abhängigkeit der definierten Entfernungsmetrik übernimmt die Berechnung dafür die spezifische Entfernungsfunktion. Dabei ist sicher zu stellen, dass zwischen den Standorten eine Entfernungsberechnung über die Funktion gegeben ist. Mit der Abgabe von Waren an den Verkaufsorten zu den Verbrauchern ist ohnehin eine Anbindung an die Infrastruktur notwendig, sodass dies als gegeben vorausgesetzt wird. In diesem Zusammenhang ist vor der Optimierung lediglich sicher zu stellen, dass keine Verkaufsorte an isolierten Positionen sind.

Zur Planung eines Verteilzentrums sind die sich potentiell ergebenden Fahrtrouten über die Entfernungsfunktion zu ermitteln. Die letztendlichen Distanzen dienen dabei zur Bewertung der Planung, um eine Auslieferung der Waren von den Verteilzentren zu den Verkaufsstandorten vorzunehmen. Über die Schnittstellenanbindung erfolgt in dem Szenario eine direkte Standort-zu-Standort Entfernungsberechnung auf Grundlage externer Daten des Kartendienstes von Google Maps gemäß Abschnitt 5.1.6. Somit lassen sich auch Informationen über die tageszeitabhängige Verkehrssituationen oder Kosten für Routen mit einbeziehen.

Bei der Platzierung von den Zwischenlagern ist auf eine Anbindung an die Infrastruktur zu achten. Da der Bau einer solchen Straßenanbindung im Zuge der Planung des Zentrums mit einplanbar ist, lässt sich in Abhängigkeit des Terrains eine freie Platzierung der Zwischenlager vornehmen. Insbesondere mit der Entwicklung von fliegenden Drohnen zum Transport verringert sich die Abhängigkeit zur Infrastruktur [28, 275].

Zudem erfolgt durch die Entfernungsfunktion eine Berücksichtigung von Standorten mit bereits vorhandener Infrastruktur, sodass diese besser bewertet werden. Somit ist die Platzierung der Zentren nicht an bestimmte Standorte bei diesem Szenario gebunden und ermöglicht in der Optimierung eine flexible Platzierung.

Durch den Kauf der Waren durch die Verbraucher an den Verkaufsstellen ergibt sich ein individuelles Profil für die verschiedenen Standorte. Die Interessen und das Kaufverhalten der Verbraucher werden dadurch an den Verkaufsstellen aufsummiert und als Bedarfsforderungen an die Zwischenlager weitergegeben. Dieses entspricht dem Nutzerprofil im vorgestellten Modellierungsansatz. Zur Differenzierung und der Entscheidung über die Wichtigkeit von Anforderungen ermöglicht das Modell mit der Zuweisung von Prioritäten zu den Nutzern eine Unterscheidung. Diese lässt sich in Abhängigkeit zum Kosten-Nutzen-Verhältnis mit den Eingangsdaten des Szenarios vornehmen.

Bei der Planung von den Standorten der Zwischenlager ist während der Optimierung zwischen den verschiedenen Verkehrsverbindungen von den Verkaufsorten zu den geplanten Lagern zu unterscheiden. Zur Berechnung und Bewertung der Entfernung zwischen den geografischen Standorten wird vorrangig die Metrik Kilometer von der schnellsten Verbindung verwendet. Zur detaillierteren Abbildung der realen Gegebenheiten lassen sich die spezifischen Eigenschaften der Straßen mit im Modell berücksichtigen. Dies ist insbesondere von Bedeutung, wenn die Wegwahl anhand mehrere Kriterien zu erfolgen hat. Neben der Entfernung sind ebenso die Fahrzeit und -kosten für ein Unternehmen relevant. Darüber hinaus ist die Häufigkeit der Belieferung als auch die potentielle Gefahr von Verzögerungen auf der Strecke zu beachten. Diese Aspekte lassen sich über

die Funktion zur Entfernungsberechnung mit einbeziehen, indem das zurückgelieferte Ergebnis einen aggregierten Wert dieser Eigenschaften darstellt. Der Wert beinhaltet in der Gesamtheit die vorab spezifizierten Eigenschaften der Infrastruktur, welche zur Bewertung der Verbindungen zu berücksichtigen sind.

Bei der Planung der Zentren lassen sich einmalige Effekte sowie dauerhafte Eigenschaften über die Nebenbedingungen mit einbeziehen. Zu den einmaligen Effekten zählen die Baukosten und Grundstückskosten. Als dauerhafte Eigenschaften sind unter anderem Unterhaltskosten sowie Lohnkosten anzuführen. Über den Modellierungsansatz von Schichten lassen sich somit mehrere Vor- und Nachteile eines Standortes mit einbeziehen. Zur Bewertung eines Szenarios werden die Werte der Schichten entsprechend akkumuliert sowie mit den Prioritäten der Nutzer und den entsprechenden Werten der Entfernungsfunktion multipliziert.

Diese beschriebenen Entitäten und Zusammenhänge ergeben ein Modell für das im Abschnitt 1.2.3 vorgestellte Beispielszenario. Es beinhaltet die priorisierten Nutzer mit deren Profil. Die vorgegebene Infrastruktur wird über eine Entfernungsfunktion eingebunden. Für die Verteilzentren besteht eine freie Platzierungsmöglichkeit zur Bereitstellung der Dienste. In Abhängigkeit der gewählten Kriterien und der Zielfunktionen erfolgt die Optimierung hinsichtlich der spezifizierten Funktionswerte mittels den Algorithmen aus dem Abschnitt 4.8 und den in dieser Arbeit entwickelten Verfahren in Abschnitt 5.3. Für das betrachtete Beispiel erfolgt zunächst eine Reduzierung der Transportwege. Das Ergebnis der Optimierung enthält die geplanten Standorte der Lager sowie die Zuweisung zu den Verkaufsstandorten. Dieses spezifische Modell lässt sich auch in anderen Modellierungswerkzeugen weiterverwenden. Abbildung 5.12 zeigt die Nutzung mittels der Software Anylogic [276].

Zur Simulation der Zwischenlagerung von Waren an den Verteilzentren besitzt jedes dieser Zentren eine vordefinierte Lagerkapazität, welche mit der Größe des Speichers angegeben ist. Da es sich bei den Waren um Stückgut handelt, wird mit der Auslieferung der Ware der Bestand im Zentrum verringert und die verfügbare Lagerkapazität zur Zwischenspeicherung freigegeben. Da die Lagerkapazität begrenzt ist, verfolgt das Zentrum eine Bestandsoptimierungsstrategie, um die notwendigen Waren im Voraus zur Verfügung zu haben. Die verschiedenen Möglichkeiten sind im Abschnitt 5.1.12 aufgeführt.

In der Simulation des Modells melden die Verkaufsstandorte deren Bedarf an benötigten Waren dem zugewiesenen Zentrum. Hierbei wird zwischen den verschiedenen Verhaltensweisen gemäß Abschnitt 5.1.12 unterschieden. Die spezifischen Parameter für die Profile der Nutzer werden mit den Eingangsdaten dem Modell übergeben. Das Zentrum übergibt die Waren entsprechend den Anforderungen an die Verkaufsstandorte ab. Anschließend wird entsprechend der spezifizierten Bestellpolitik vorgegangen, um den Bestand des Lagers wieder aufzufüllen. Die Ergebnisse der Simulation werden anschließend in die Optimierung mit einbezogen, sodass sich die verschiedenen Verhaltensweisen der Nutzer mit auswirken. Dazu erfolgt eine Korrelation der Nutzerprofile und eine angepasste Zuweisung zu den Verteilzentren entsprechend der Zielfunktionen aus Abschnitt 5.1.9. Durch den iterativen Optimierungsprozess werden die verschiedenen Zielkriterien für eine ganzheitliche Lösung berücksichtigt.

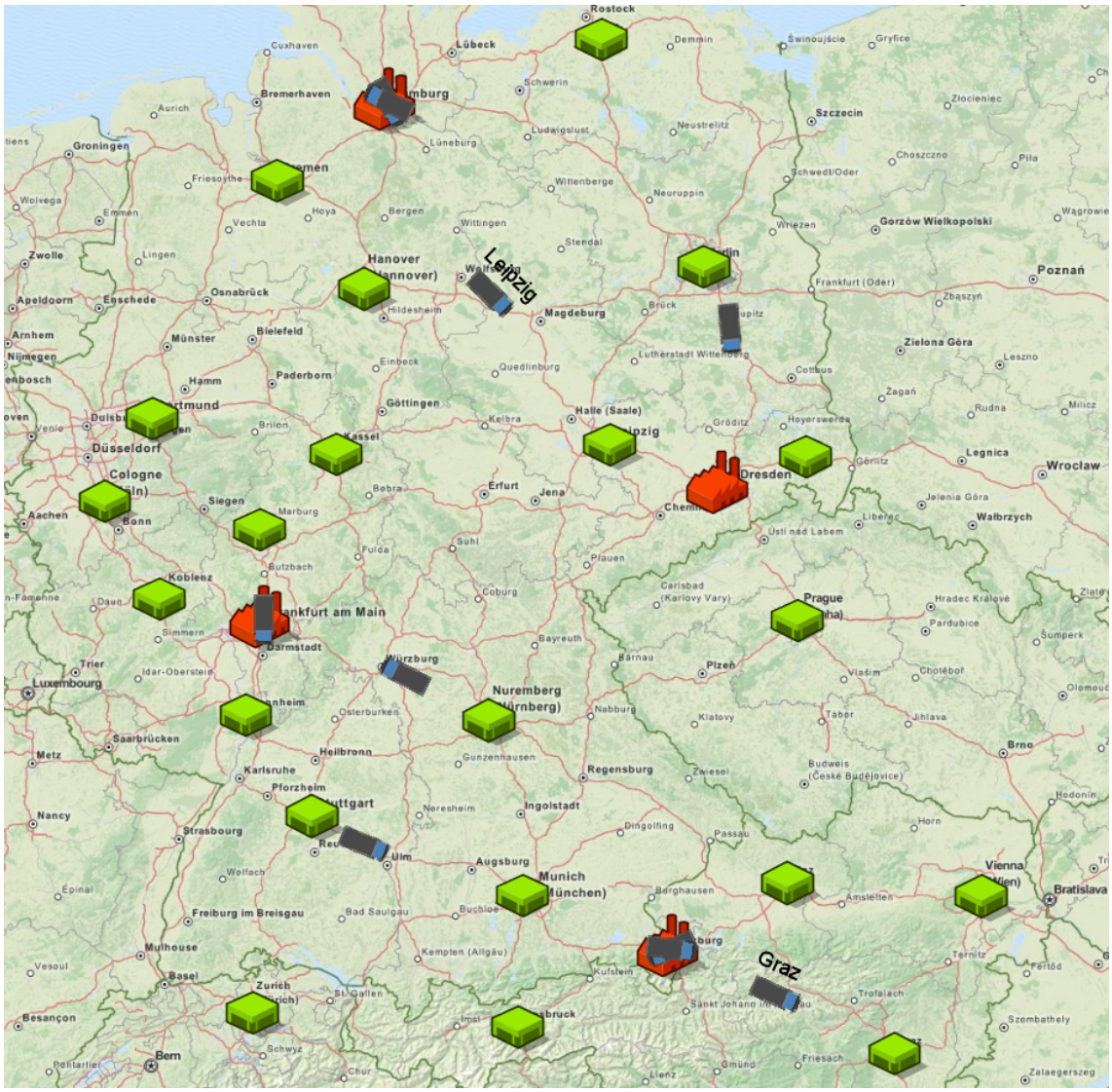


Abbildung 5.12: Anwendung des generischen Modells auf ein spezifisches Szenarios der Logistik mit der Visualisierung der Ergebnisse in dem Simulationswerkzeug Anylogic [276]

5.2.2 Spezifisches Modell für Rechnernetze

In diesem Abschnitt wird der generische Modellierungsansatz auf den Bereich von Rechnernetzen angewendet und die Umsetzung erläutert. Zur Erstellung eines spezifischen Modells für CDN werden zunächst die spezifischen Eigenschaften des Anwendungsgebietes näher analysiert. Die Übertragung der Daten von einem Server zu einem Nutzer erfolgt über die vorhandene Netz-Infrastruktur. Die Server werden über dedizierte Netzknoten an diese Infrastruktur angeschlossen. Daher ist das Modell direkt von der zugrundeliegenden Netz-Infrastruktur abhängig.

Die globale Internet-Infrastruktur besteht aus Netzen mit unterschiedlicher administrativer Verwaltung, welche zusammengeschaltet werden. In Abhängigkeit der Größe und transnationalen Verbindungen lassen sich die Netze der Internet-Dienstanbieter in drei Klassen einteilen. Diese werden als Tier Level bezeichnet. Tier 1 Netze sind Weitverkehrsverbindungen mit autonomen Systemen auf oberster Ebene. Die Netze

sind mindestens national weit verbreitet und besitzen globale Verbindungen zu anderen Providern. Tier 2 Provider erkaufen sich Konnektivität zu Tier 1 Netzen und verkaufen diese zu Tier 3 Provider. Darüber hinaus wird Transit Funktionalität bereitgestellt. Auf der untersten Ebene der Topology befinden sich die Tier 3 Netze von lokalen Providern. Diese besitzen keine eigenen Transit Fähigkeiten und bieten vorwiegend Netzzugang für Endanwender.

Für eine effiziente Datenübertragung der vom CDN angebotenen Dienste, sind die zugehörigen Spiegel-Server an entsprechend strategische Knotenpunkte anzubinden. Ein Spiegel-Server beantwortet eine hohe Anzahl von Nutzeranfragen von einer örtlich weit verteilten Menge an Nutzern. Zur Bewältigung der Anfragelast und der damit zu erwartenden Netzauslastung sind die Server eines CDN an Netzverbindungen mit hoher Bandbreite und der Möglichkeit zur Transitdatenübertragung zu verbinden. Die Top-Level Infrastrukturen der Tier 1 oder 2 Topologie bieten einen solchen Netzzugang mit hoher Leistungsfähigkeit, wodurch deren Netzknoten sich zur Standortwahl anbieten. Dadurch lassen sich die Möglichkeiten bei der Planungsgrundlage für die Standortauswahl auf die existierenden Knotenpunkte dieser nicht hierarchischen Topologien begrenzen. Darüber hinaus wäre es unvorteilhaft einen Server an einen Standort zu platzieren, wobei keine ausreichende Infrastruktur vorhanden ist. Ebenfalls ist die Platzierung in der Mitte einer Datenleitung ohne direkte Anschlussmöglichkeit unpraktikabel. Die Länge der Datenleitung zur Anbindung eines Servers an einen Netzknoten wird im Modell bei der Standortauswahl vernachlässigt, weil diese unabhängig von der Platzierungsentscheidung bei jeder Serveranbindung entsteht. Der Einfluss von solch kurzen Längen ist vernachlässigbar in Relation zum gesamten Netzpfad. Daher entspricht die Planung von Servern eines CDN der Platzierung an dedizierten Knotenpunkten einer Infrastruktur, die im Modell als Graph abgebildet ist.

Die Modellierung von Nutzern des CDN erfolgt über die Eingabeinformationen des zu Grunde liegenden Szenarios. Dies beinhaltet die geographischen Positionen der Nutzer sowie deren Anbindung an das Netz über einen Internetgateway entsprechend der vorhandenen Infrastruktur. Der Zugang zu Transitnetzen erfolgt bei jedem Nutzer in Abhängigkeit deren Netzanbindung über vordefinierte Netzknoten. Da dieser Verbindungspfad beim Zugriff auf Weitverkehrsnetze auf Grund der Infrastruktur fest vorgegeben ist und sich als konstant annehmen lässt, wird jeder Nutzer dem zugehörigen Zugangsknoten im Weitverkehrsnetz zugewiesen [52]. Mehrere, einem Netzknoten zugeordnete Nutzer bilden somit eine Gruppe. Diese Gruppe von Nutzern wird im Modell zu dessen gemeinsamen Netzknoten und dessen Standort im Weitverkehrsnetz abstrahiert. Im generischen Modellierungsansatz bilden diese Knoten die Nutzer, welche Anfragen nach Daten ans CDN stellen.

Durch die Abstraktion und Gruppierung von mehreren Nutzern zu einem zugehörigen Netzknoten, wird die Anzahl von Eingangsdaten in dem Modell reduziert und die Komplexität gemindert. Mit der Gruppierung der Nutzer zu dem Standort des Netzzugangsknotens werden auch die Eigenschaften der Nutzer aggregiert. Für das Modell ist das Abstraktionslevel ausreichend präzise, da alle Anfragen der Nutzer den entsprechenden Netzknoten passieren und somit die Standortwahl der Server nicht beeinflussen. Im Modell werden die durch das CDN angebotenen Daten als

Dienste aufgefasst. Mit der Gruppierung von den Nutzer kombinieren sich auch die Bedürfnisse und das Verhalten, welche im Modell zu einem gemeinsamen Profil des Knotenpunktes aggregiert wird. Um diese kombinierten Eigenschaften von unterschiedlichen, abstrahierten Nutzergruppen gegeneinander zu differenzieren, wird dem entsprechenden Netzknoten eine Priorität zugewiesen. Diese ist abhängig von der Anzahl der Nutzer in der Gruppe, deren Wichtigkeit für den CDN Provider, der Vergütung des Dienstes und deren Anfrageverhalten sowie weiterer beeinflussenden Eigenschaften.

Bei der Berechnung und der Optimierung eines zu bevorzugenden Serverstandortes ist zwischen den verschiedenen Verbindungen einer Nutzergruppe zu den geplanten Serverstandorten zu unterscheiden. Zur Berechnung der Entfernung zwischen einzelnen Netzknoten wird vorrangig die Metrik der Hops verwendet. Bei einem Hop handelt es sich um einen Netzknoten, den ein Datenpaket bei einer Kommunikation passiert. Ein Netzknoten in der Entfernung von null Hops entspricht dem eigenen Knoten. Die Verzögerung bei der Datenübertragung wird hauptsächlich durch die Verarbeitungszeit durch den Knoten beeinflusst als durch die Übertragungszeit über ein spezifisches Medium. Demzufolge resultiert die Entfernung zwischen zwei beliebigen Knoten von einer Infrastruktur in der Anzahl der Hops. Die Berechnung des kürzesten Pfades erfolgt in dem Szenario mittels dem Dijkstra Algorithmus.

Zur genaueren Abbildung der Infrastruktur und des Sachverhaltes werden die spezifischen Eigenschaften eines Datenlinks mit im Modell abgebildet. Dies ist von Bedeutung, wenn es mehrere Pfade zu den geplanten Servern von einem Netzknoten aus mit der gleichen Hop Anzahl gibt. Dementsprechend wird ein genauerer Funktionswert zur Evaluierung benötigt, welcher unter anderem die spezifischen Verbindungseigenschaften maximale Bandbreite, durchschnittliche Auslastung und Verzögerung sowie Redundanz mit einbezieht. Diese Informationen müssen bei der Modellerstellung vorzugsweise vom ISP als Betreiber der Infrastruktur bereitgestellt werden. Abbildung 5.13 zeigt dafür ein Beispiel des ISP Dial Telecom [277].

Bei genauerer Betrachtung von gesamten Verbindungspfaden durch eine Infrastruktur weisen die einzelnen Teilverbindungen unterschiedliche Eigenschaften auf. Bisher sind keine multikriteriellen Wegfindungsalgorithmen definiert, welche direkt mit mehreren Verbindungseigenschaften arbeiten. Daher ist ein direkter Vergleich der Verbindungsgüte über einen gesamten Verbindungspfad hinweg nur bedingt möglich. Ein Beispiel des Problems ist in der Abbildung 5.14 dargestellt.

Zum Vergleich von gesamten Verbindungspfaden mit anderen Pfaden ist ein einzelner Vergleichswert für eine Ende-zu-Ende-Verbindung notwendig. Dieser ist entsprechend der fokussierten QoS und dem Zielkriterium zu spezifizieren. Da einzelne Teilverbindungen auch Bestandteil anderer Verbindungspfade sind, ist den einzelnen Teilverbindungen ein spezifischer Wert zuzuweisen. Der Vergleichswert eines gesamten Verbindungspfades setzt sich dementsprechend aus den Wichtungswerten der Teilverbindungen zusammen, indem diese aufsummiert werden. Zur Abstraktion zu einem Zielfunktionswert sind zunächst die Verbindungseigenschaften entsprechend der Zielfunktion zu einem Wichtungswert zu kombinieren. Dieser Wert wird mit dem definierten Wert eines Hops kombiniert und das Ergebnis als eine virtuelle Verbindungslänge aufgefasst, welche größer gleich eins ist. Eine Verbindung mit hoher Leistungsfähigkeit hat eine virtuelle Länge von eins, was den

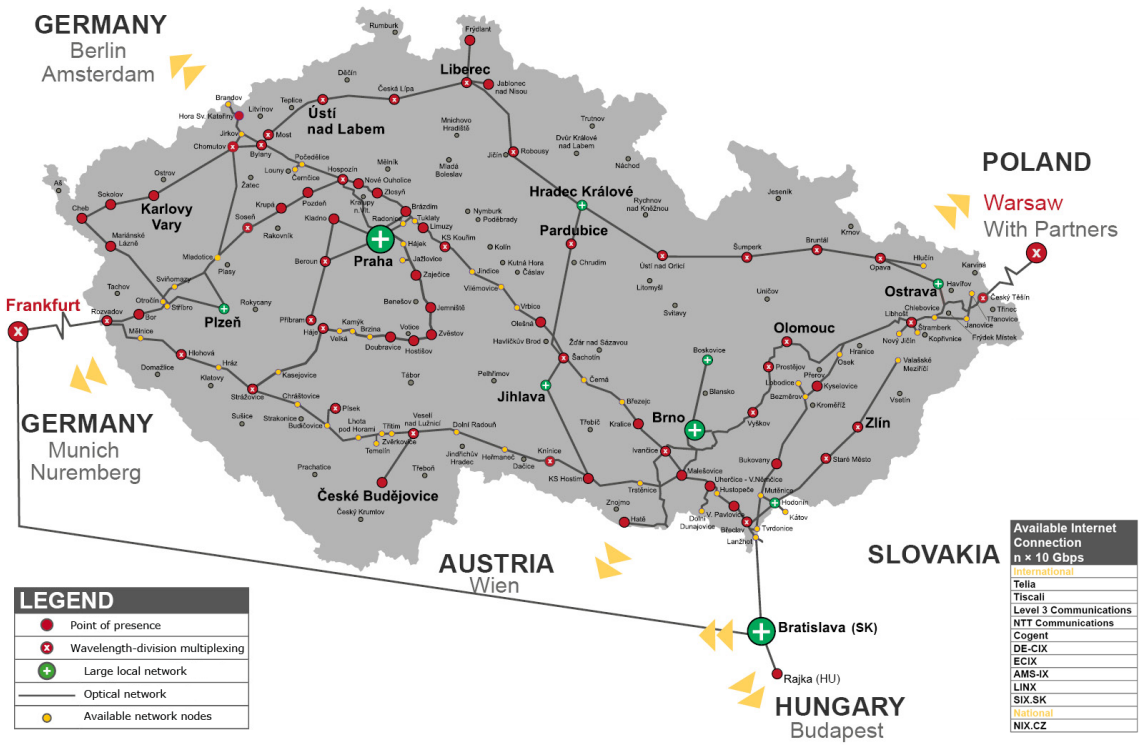


Abbildung 5.13: Backbone Infrastruktur der Dial Telecom in der Tschechischen Republik mit Anbindung zu weiteren Ländern (Vgl. [277])

bestmöglichen Wert repräsentiert. Verbindungen mit geringerer Leistungsfähigkeit wird ein entsprechend proportional größerer Wichtungswert zugewiesen, was einer längeren virtuellen Strecke entspricht. Durch die Abstraktion der Verbindungseigenschaften auf eine virtuelle Länge ermöglicht es für die Wegfindung weiterhin die Verfahren zur Berechnung kürzester Pfade in Graphen gemäß Abschnitt 5.1.6 zu verwenden.

Somit lässt sich zur Unterscheidung von Verbindungen von einem Nutzer zu mehreren Servern die akkumulierte, virtuelle Pfadlänge verwenden. Dies ermöglicht es verschiedene Verbindungspfade zu vergleichen und damit eine effiziente Zuweisung von Nutzern zu Servern vorzunehmen. Das beschriebene Modell entspricht der Anwendung des generischen Modellierungsansatzes auf das Beispielszenario aus dem Abschnitt 1.2.4. Dieses beinhaltet die priorisierten Netzknoten von einer Gruppe von Nutzern in der vorgegebenen Infrastruktur und die Platzierungsmöglichkeiten für Server zur Verfügungsstellung des Dienstes eines Anbieters. Um den Einfluss der Nutzerpriorität mit bei der Evaluation zu berücksichtigen, wird der Wert der virtuellen Pfadlänge mit dem Wert der Priorität des entsprechenden Knotens multipliziert. In Abhängigkeit des gewählten Kriteriums und der Zielfunktion erfolgt die Optimierung über die spezifizierten Zielfunktionswerte mittels den Algorithmen aus den Abschnitten 4.8 und 5.3. Für die betrachteten Beispiele erfolgt vorerst eine Fokussierung auf eine Reduzierung der Netzlast mit kurzen Übertragungswegen. Das Ergebnis der Optimierung enthält die geplanten Standorte von Servern sowie die Zuweisung der Nutzer.

Zur Simulation des Speicherverhaltens eines Servers besitzt jeder Spiegel-Server eine vordefinierte Menge an Speicherplatz. Ein einmal zwischengespeicherter Inhalt lässt sich

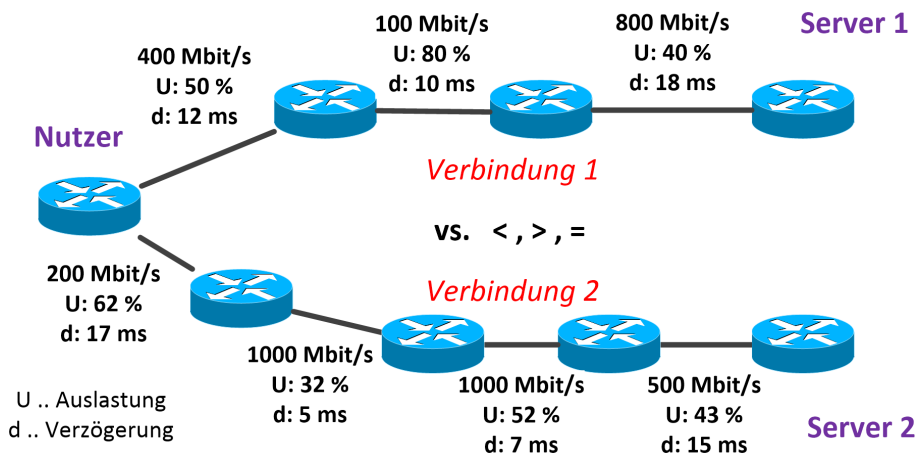


Abbildung 5.14: Problem des Vergleichs verschiedener Netzverbindungen von einem Nutzer zu einem Server

dabei beliebig oft abrufen. Da die verfügbare Menge an Speicherplatz wesentlich kleiner ist als die in der Gesamtheit angebotenen Daten, benötigt der Server eine Strategie zur Verwaltung der angebotenen Daten. Im Modell lassen sich dafür die verschiedenen Möglichkeiten aus Abschnitt 5.1.12 verwenden.

Während der Simulation wird jeder Netzknoten wiederholend betrachtet, ob von diesem aus eine Anfrage an einen definierten Server gesendet wird. Dabei bestehen die verschiedenen Verhaltensmöglichkeiten von Nutzern aus Abschnitt 5.1.12. Die von einem Nutzer angefragten Daten sind in dem Profil der Netzknoten beschrieben. Hierbei bieten sich als Profil zuvor aufgezeichnete Traces an, welche durch das Szenario dem Modell zu übergeben sind. Das Ergebnis der Simulation wirkt sich anschließend auf die Optimierung aus. Hierbei wird das Nutzerverhalten über die Korrelation der Anfrageprofile mit einbezogen, indem eine angepasste Zuweisung der Nutzer zu den Servern erfolgt. Dabei bieten sich die verschiedenen Korrelationsfunktionen gemäß Abschnitt 5.1.9 an. Die Korrelation eines Nutzerprofiles mit dem aggregierten Profil des Zentrums bestimmt das Gütemaße, in wie weit das Verhalten eines Nutzer mit dem vom Server übereinstimmt. Durch den sequentiellen Optimierungsansatz wird eine ganzheitliche Lösung erreicht.

5.2.3 Spezifisches Modell für Geolokalisation

Ein weiteres Anwendungsgebiet für das generische Modell ist die Geolokalisation. Der Begriff der Geolokalisation umschreibt den Prozess zur Bestimmung des physischen Standortes, z. B. anhand der Geokoordinaten Breitengrad und Längengrad, von einer logischen Adresse. In dem spezifischen Anwendungsgebiet von Computernetzen erfolgt die Bestimmung des geographischen Standortes von einem Netzgerät anhand seiner logischen Internet Protokoll (IP) Adresse. Dies wird als IP Geolokalisation bezeichnet [278–281]. Dafür wird in diesem Kapitel zusätzlich ein spezifisches Modell für die Geolokalisation entwickelt und der Zusammenhang mit dem generischen Ansatz dargestellt. Dies betrifft insbesondere die Bestimmung von den dafür notwendigen zentralen Messpunkten in Rechnernetzen.

Immer mehr Dienste berücksichtigen den geographischen Standort der Nutzer, von wo aus der Zugriff erfolgt. Anwendungsfälle sind unter anderem CDN in Bezug auf optimierte Lastverteilung und Verkehrsmanagement [20, 282, 283], personalisierte Werbung [284] und Cyberkriegsführung [285]. Vor allem für Strafverfolgungsbehörden ist eine präzise Standortbestimmung zwingend erforderlich, um einen Angreifer zurück zu verfolgen.

Arten von IP Geolokalisation

Die verschiedenen Strategien für IP Geolokalisation sind in die Klassen abbildungsbasierte, messungsbasierte und hybride Verfahren eingeteilt [286].

Abbildungsbasierte Verfahren: Die abbildungsbasierten Ansätze werden auch als passive Verfahren bezeichnet, da hierbei keine Interaktion mit dem Netzgerät und der entsprechenden IP-Adresse erfolgt. Der geografische Standort wird unter anderem über öffentlich zugängliche Datenbanken wie den fünf *Regional Internet Registries* abgefragt, welche von der *Internet Assigned Numbers Authority* (IANA) zur Verwaltung des lokalen IP-Adressraums delegiert sind [278, 280, 286, 287]. Weitere Möglichkeiten sind Hinweise über das *Domain Name System* (DNS) mit den *Location Resource Record* [288, 289] zu erhalten und Abfragen über *Whois-Lookup* [290] oder kommerzielle Datenbanken von Geodiensten [291] zu tätigen. Darüber hinaus gibt es zur passiven Geolokalisation die Methoden von *NetGeo* [292] und *Structon* [293] sowie die *IP2Geo* Suite [278], welche aus *GeoCluster* und *GeoTrack* besteht. Hierbei wird auf den Standort zu einer IP-Adresse in Anlehnung anderer Netzgeräte und Routing Informationen geschlossen, bei denen der Standort bekannt ist und Verbindungsinformationen gemäß dem *Border Gateway Protocols* erhältlich sind. Eine weitere Alternative besteht in der Abfrage vom *Fully Qualified Domain Name* (FQDN) der Router oder anderer Netzkomponenten, welche auf dem Pfad zur Ziel-IP-Adresse liegen. Bei passiven Methoden ist zu berücksichtigen, dass diese insbesondere auf Grund der dynamischen IP-Zuweisung bei privaten Nutzern nur grobe und unvollständige Informationen liefern [280, 286, 294].

Messungsbasierte Verfahren: Die Klasse der messungsbasierten Verfahren beruht auf einer aktiven Interaktion mit der entsprechenden Netzkomponente einer IP-Adresse. Hierbei werden dem System mit der zugehörigen IP-Adresse Datenpakete über das Netz gesendet und die Zeitspanne für die benötigte Antwort erfasst. Diese Strategien basieren auf der nachgewiesenen Annahme einer bestehenden Korrelation zwischen Netzlatenz und geographischer Entfernung [29]. Diese Anfragen erfolgen von sogenannten Vantage Points (VP) aus. Ein VP repräsentiert eine Netzkomponente mit Messinfrastruktur, dessen geografischer Standort bekannt ist [295]. Die gesendeten Datenpakete bestehen aus *Echo Requests* gemäß Ping und Traceroute Anfragen. Zudem werden in der Geolokalisation Landmarks verwendet, was reaktive Netzkomponenten mit IP-Adressen sind und deren Standort bekannt ist. Technische Lösungsansätze auf Grundlage von reinen Messergebnissen der Verzögerungen sind *Shortest Ping* [296] und *GeoPing* [278]. Bei *Shortest Ping* wird die Ziel-Adresse auf einen Landmark mit möglichst gleich geringer

Latenz abgebildet. *GeoPing* ist teil von *IP2Geo* und bestimmt den gemittelten Standort über Entfernungsvektoren, welche auf den Verzögerungen von den Anfragen beruhen.

Weiterentwickelte Methoden auf Grundlage von aktiven Messungen nutzen zusätzliche Informationsquellen, um Messungenauigkeiten und Toleranzen auszugleichen. Das *Constraint-Based Geolocation* (CBG) [297] bestimmt die geographische Lage durch Überlagerung von Entfernungsbereichen, wobei Grenzen definiert werden. Dadurch entsteht ein kontinuierlicher Lösungsbereich, anstatt ein konkreter Ort mit Geokoordinaten angegeben wird. Die *Topology-Based Geolocation* (TBG) [296] ist eine Erweiterung von CBG und ermittelt über Topologie-Messungen den Standort des Ziels als auch die Position von Zwischenrouter zur höheren Zuverlässigkeit.

Ein Beispiel für den Prozess der aktiven, messbasierten IP Geolokalisation ist in der Abbildung 5.15 dargestellt. Insbesondere im Bezug auf die Aktualität bei sich ändernden IP-Adressen von Netzkomponenten und dem damit zusammenhängenden Standort haben sich aktive Geolokalisationsverfahren als präziser erwiesen [294–296].

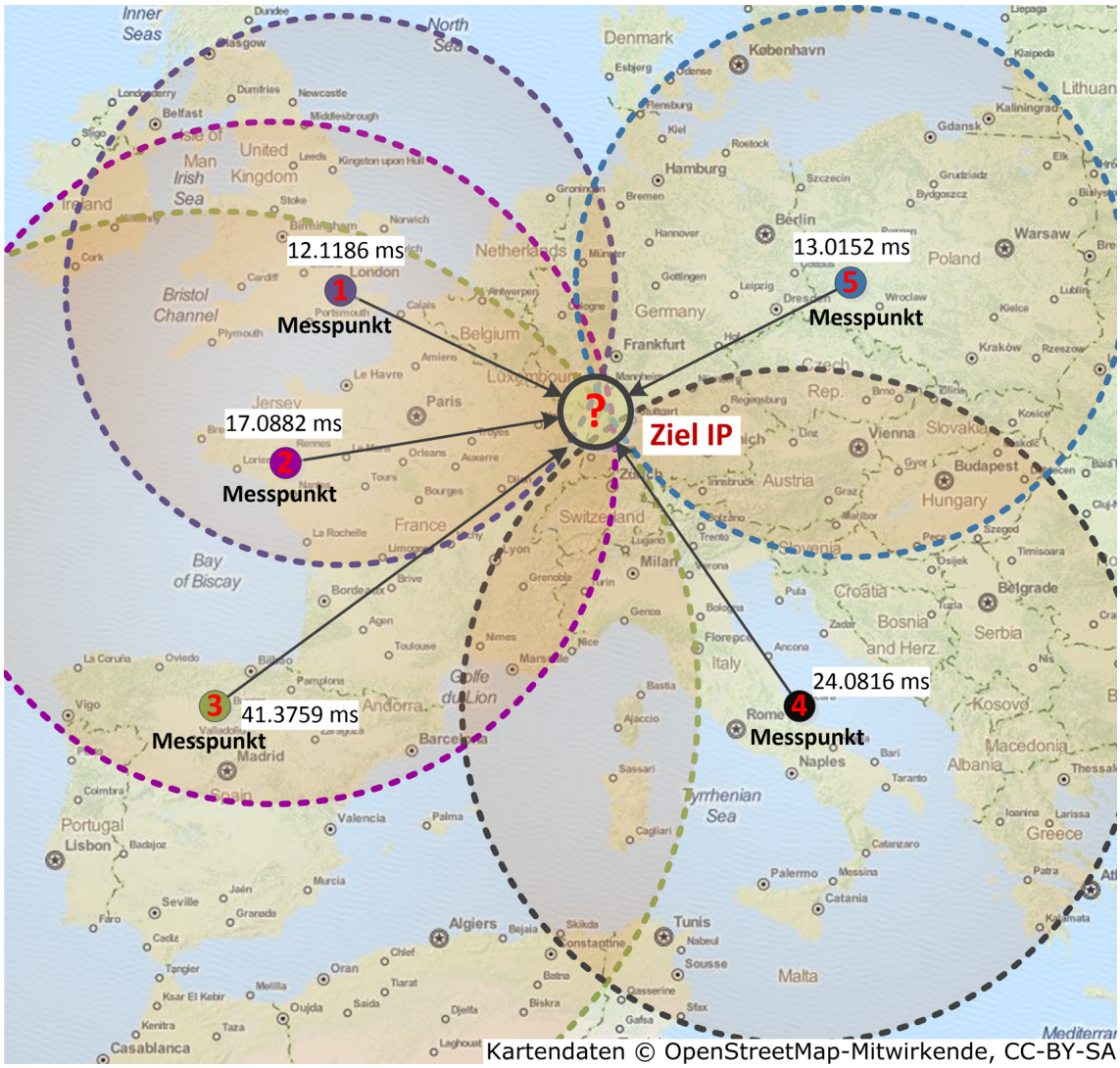


Abbildung 5.15: Beispiel für eine IP Geolokalisation über aktive Messungen

Hybride Verfahren: Hybride Ansätze nutzen die zuvor genannten Verfahren und wichten entsprechend die Ergebnisse, um diese zu verifizieren als auch Nachteile auszugleichen [286, 298]. Einer der bedeutendsten Vertreter ist das Framework *Octant* [284]. Der Standort wird durch den Einsatz von überlagernden *Bézierkurven* eingegrenzt, wobei weitere positiven als auch negative Nebenbedingungen herangezogen werden. Durch das komplexe Verfahren lässt sich die geographische Region genauer eingrenzen als bei CBG [295]. Das modulare Design erlaubt es zusätzliche Einschränkungen wie demographische Daten zu verwenden, um die Genauigkeit zu verbessern [299]. Weitere Beispiele für hybride Ansätze sind *HawkEyes* [279], *Spotter* [280] und *POSIT* [281]. Diese betrachten weitere Nebenbedingungen in der Akkumulation von Verzögerungen und Entfernungen. Dabei werden unter anderem urbane und nicht urbanen Regionen einbezogen, um den Standort weiter einzugrenzen, wobei keine genaue Position angegeben wird. Diese Ansätze berücksichtigen zudem keine realen Umstände in Rechnernetzen.

Problem der Identifizierung von Vantage Point

Alle messungsbasierten Verfahren verwenden VPs für die Analysen, welche aktiv Anfragen an die Ziel-IP-Adresse senden und damit Latenzmessungen durchführen. Die Identifizierung von optimierten Standorten für Messpunkte in einer Netzinfrastruktur sowie die notwendige Anzahl solcher VPs stellt dabei ein NP-schweres Problem dar. Mehr VPs erhöhen voraussichtlich die Zuverlässigkeit der Messergebnisse, aber verursachen auch erhöhte Betriebskosten. Dies ist vergleichbar mit der Problemstellung der Platzierung von Zentren in Kapitel 2.3.6. Die Literatur zur IP Geolokalisation bietet keine umfassenden Informationen über die optimale Auswahl der VPs und wie mit dem Problem umzugehen ist. Lediglich die Forschungsarbeit [29] beschreibt ein Verfahren, welches auf Basis eines ILP-Modells die Platzierung von VPs in einem Netz vornimmt. Dieser Ansatz entspricht dem in Kapitel 4.8.6 vorgestellten Verfahren. Auf Grund der Berechnungskomplexität und der Integritätslücke von zwei ist der beschriebene ILP Ansatz für große Szenarien nicht praktikabel einsetzbar und beschränkt sich auf spezifische Regionen. Darüber hinaus werden bei gegenwärtigen Modellen der Geolokalisation euklidische und orthodrome Metriken als approximierte Entfernungen zwischen zwei Standorten verwendet anstelle der zu Grunde liegenden Netzinfrastruktur, was zu einer ungenauen Modellierung des Problems führt. Dieser Zusammenhang wird durch die Problembeschreibung im Kapitel 2.1.1 thematisiert und im generischen Modell im Kapitel 5.1.6 umgesetzt.

Zusammenhang mit dem generischen Modell

Im Bezug zum generischen Modell werden dessen Fähigkeiten für einen neuen Ansatz zur messungsbasierten Geolokalisation eingesetzt. Der Ansatz nutzt die moderate Korrelation zwischen Netzverzögerung und geografischen Abstand aus. Die Genauigkeit ist zu einem wesentlichen Bestandteil durch die Auswahl der Messpunkte und die mathematische Modellierung beeinflusst. Der Ansatz verwendet das generische Modell zur Berechnung optimierter Standorte für die VPs bei aktiven Messungen. Mit einer verbesserten

Platzierung und Verteilung der Messpunkte ergeben sich kurze Messstrecken, was zu weniger Störeinflüssen und damit genaueren Ergebnissen führt. Dabei lassen sich verschiedene Optimierungsalgorithmen zur Identifizierung der VPs einsetzen. Insbesondere betrifft dies den neu entwickelten Platzierungsalgorithmus. Weiterhin eignet sich der Algorithmus zur Bestimmung des gesuchten Standortes aus den erhaltenen Messergebnissen durch Clustering. Zur Erreichung genauer Ergebnisse sind reale Einflüsse in den Netzinfrastrukturen zu berücksichtigen. Hierbei kommt insbesondere die Flexibilität der Entfernungsfunktion des generischen Modells zum Tragen.

Im Hinblick auf die Anwendung des Systems bei CDN und der Bearbeitung von Nutzeranfragen lassen sich Anfragen durch einen Steuerungsmechanismus an einen der Server im CDN Verbund weiterleiten. Entsprechend der Verfügbarkeit der Daten auf einem Server, der Auslastung des Servers und dem Standort ist eine passende Auswahl zu treffen. Insbesondere der Standort-Aspekt von dem Anfragenden lässt sich mittels Standortbestimmung über Geolokalisation in den Steuerungsmechanismus mit einbeziehen. Die Zuweisung von Anfragen zu einem bestimmten Server wird in diesem Fall mittels Geolokalisation gesteuert. Hierzu wird die IP-Adresse des Nutzers aus der Anfrage extrahiert und an die Server des CDN verteilt. Diese führen im Verbund eine IP Geolokalisation des anfragenden Netzknotens durch indem diese Messungen zu der IP-Adresse vornehmen. Alle Messergebnisse werden anschließend bei einem Server gesammelt, welcher mittels Berechnungen aus den Messdaten den Standort bestimmt. Somit wird die Server-Infrastruktur vom CDN gleichzeitig zur Geolokalisation genutzt. Anschließend wird die Anfrage des Nutzers an den Server weitergeleitet, welcher unter Beachtung der Netzinfrastruktur vorteilhaft gelegen ist. Somit wird eine Reduzierung der Netzlast erreicht und die QoS für den Nutzer erhöht.

Modell für die Geolokalisation

Neben dem generischen Modellierungsansatz wird im Folgenden ein weiteres Modell für die Geolokalisation aufgestellt, dass spezifisch für den Bereich der aktiven Geolokalisation ist und Teile des generischen Modells mit nutzt. Eine Übersicht der einzelnen Prozessschritte ist in der Abbildung 5.16 dargestellt.

Das Modell zur Geolokalisation verwendet eine vordefinierte Anzahl von fest platzierten VPs, um Messungen zur Ziel-IP-Adresse durchzuführen. Dazu sind diese Standorte der VPs unter Beachtung der Netzinfrastruktur zu bestimmen. Das Ziel besteht dabei in der Identifizierung von zentralen Knotenpunkten sowie in der Reduktion der Länge von Netzpfaden von einem VP zu dem umliegenden Bereich. Die zentralen Knoten ermöglichen somit Messungen in verschiedene Teilnetze mit reduzierten Netzpfaden. Dadurch wird die Wahrscheinlichkeit von Störeinflüssen auf dem Pfad minimiert und somit die Variabilität als auch die Abweichung der Messergebnisse reduziert, was dessen Qualität erhöht. Die ausgewählten VPs ermitteln die *Round-Trip-Time (RTT)* und die Anzahl der Hops zur Ziel-IP-Adresse in Abhängigkeit des eingesetzten Übertragungsprotokolls und deren Parameter.

Die Messergebnisse werden für die zeitliche Korrelation verwendet, um die Entfernung von jedem VP zum geographischen Standort des Ziels zu bestimmen. Durch Lateration

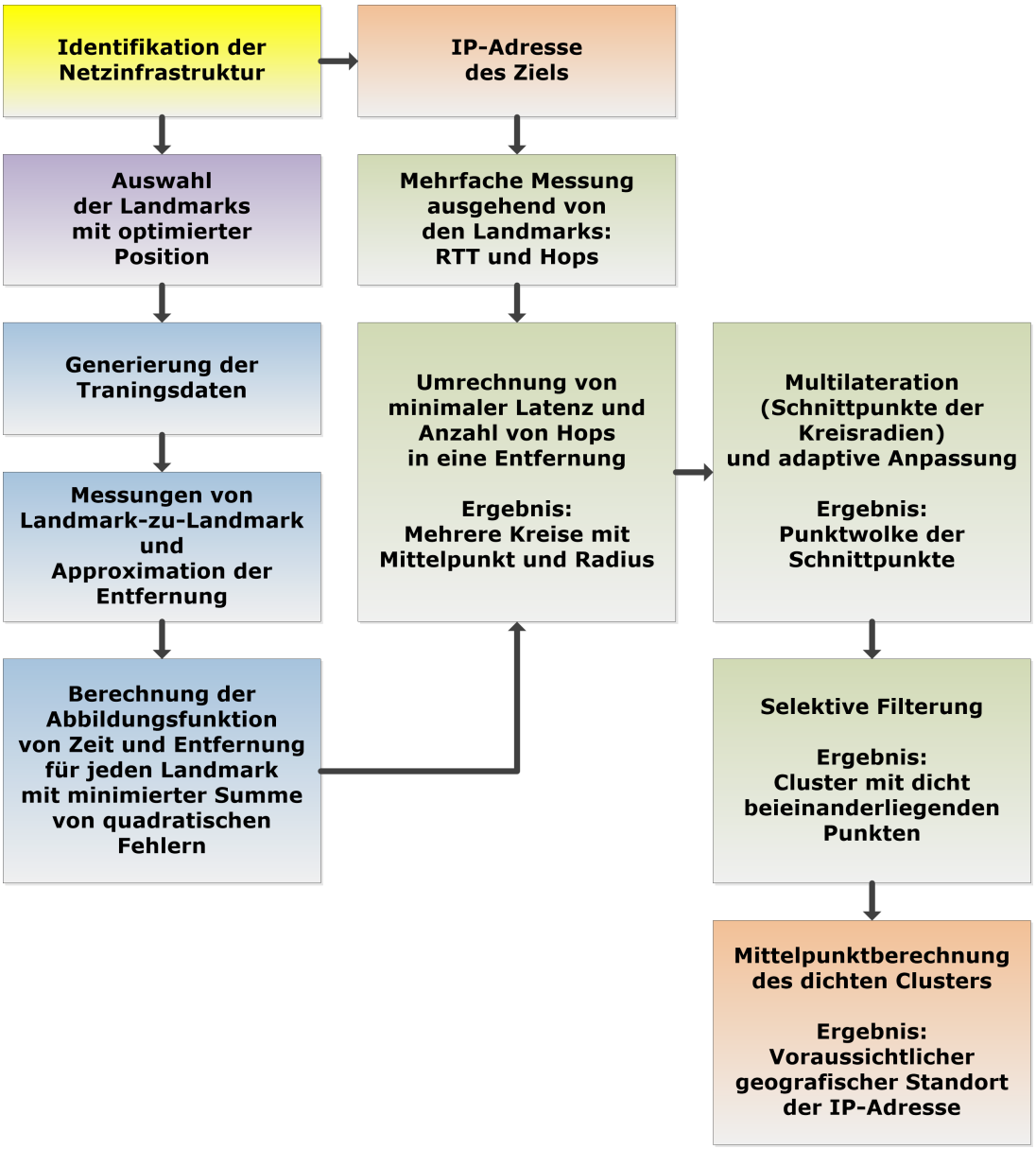


Abbildung 5.16: Prozessschritte des Modells zur Geolokalisation

von den bekannten Standorten und der Entfernung lässt sich das Ziel lokalisieren. Durch mehrfache Berechnungen unterschiedlicher VPs erfolgt somit eine Multilateration, um Messungenauigkeiten zu kompensieren. Hierfür bieten sich unter anderem hyperbolische und zirkuläre Algorithmen [300] an. Zur Steigerung der Genauigkeit und als Überprüfung ist die Netzkomponente einen oder mehrere Hops vorher ebenfalls zu lokalisieren, da eine örtliche Nähe anzunehmen ist [278]. In Einklang mit Nachverfolgungstechniken für gesendete Datenpakete lässt sich der Standort verifizieren [301]. Unter der Annahme, dass die einzelnen Hops jeweils zueinander eine geographische Nähe aufweisen, wird die Identifizierung von speziellen Netzkonfigurationen möglich, wie z. B. Proxy und VPN Verbindungen sowie MPLS Tunnel und Anonymisierungstechniken.

Die Umformung der Latenzzeit in eine Entfernung erfolgt über eine approximierte Funktion, da der direkte Zusammenhang auf Grund von Störeinflüssen komplex ist. Die

Abbildungsfunktion wird durch entsprechende Trainingsdaten ermittelt. In Kombination mit den Messergebnissen erfolgt über mehrere Berechnungen die Multilateration ausgehend von den verschiedenen VPs. Mit Messungen aus verschiedenen Richtungen und Teilnetzen werden Unter- und Überschätzungen ausgeglichen. Darüber hinaus werden die Messdaten und Teilergebnisse während den Berechnungen gezielt und automatisiert gefiltert, um die Genauigkeit des Modells weiter zu erhöhen, Störfaktoren zu minimieren und ein zuverlässiges System zu erhalten.

Das Modell zur IP Geolokalisation wird im Folgenden mit den Details zu jedem Prozessschritt beschrieben. Insbesondere ist der Einsatz des generischen Modells und die in dieser Arbeit entwickelten Optimierungsverfahren hervorgehoben, um die Vielfältigkeit des Anwendungsgebietes zu verdeutlichen. Die einzelnen Teile des Modells, Auswahl der VPs, Transformation der Latenzzeit in eine Entfernung sowie die Standortbestimmung werden in einzelnen Abschnitten erläutert.

Auswahl der Vantage Points zur Messung

Zuerst erfolgt die Identifizierung der zu Grunde liegenden Netzinfrastruktur und die Platzierung der VPs, bevor eine unbekannte IP-Adresse in dem Netz sich bestimmen lässt. Zur Auswahl der Messpunkte muss die Topology des Kernnetzes vorab bekannt sein indem Informationen von ISPs oder öffentlichen Quellen genutzt werden. In dem betrachten Szenario wird das gemeinsame Backbone Netz der ISPs in Europa verwendet [19], welches über 1100 Knoten und 1400 Verbindungen aufweist. Dieses ist in der Abbildung 5.17 dargestellt.

Bei der Platzierung besteht das Ziel die maximale Distanz zur umgebenden Netzwerktopologie eines einzelnen VP zu minimieren. Da der Standort des Ziels vorab unbekannt ist, sind die VPs gleichmäßig zu verteilen, um aus verschiedenen Richtungen unter Verwendung unterschiedlicher Netzpfade zu messen. Mit einer optimierten Auswahl befindet sich anschließend ein einzelnes VP näher an dem Ziel und die VPs stellen zentrale Knotenpunkte in der Topologie dar. Um so dichter ein VP an dem Ziel ist, desto weniger Störquellen wirken sich auf die Messergebnisse aus, was zu einer Reduzierung der Abweichungen und Varianzen führt. Im Zusammenhang mit dem generischen Modell besteht dementsprechend das Optimierungskriterium in der Minimierung der maximalen Entfernung eines Netzknotens zu seinem VP. Zur Auswahl optimierter Positionen von VP lassen sich unter anderem die verschiedene Verfahren in Kapitel 4.8 einsetzen.

In Abhängigkeit der zur Verfügung stehenden VPs muss eine Abbildung auf diese von den durch die Optimierung identifizierten Messpunkte erfolgen. Dazu werden die berechneten Standorte zu den nahe gelegenen, verfügbaren VPs zugeordnet. Eine geringe Abweichung ist im Verhältnis zu den Entfernungen in der Netztopologie vernachlässigbar, zumal ein dichtes Netz möglicher Standorte besteht. Die ausgewählten Positionen werden als VPs verwendet, um Pakete an das Ziel zu senden und die Kommunikation zwischen den VPs zu messen.

Abbildung 5.17: Backbone Netz von Europa (Netz: Interconnect [19])

Analyse der Einflüsse auf die Messergebnisse

Die ermittelten VPs führen Messungen zum Ziel und zu den anderen VPs durch, um die RTT und die Anzahl der Hops zu bestimmen. Die Messungen werden dabei mehrfach durchgeführt, um dadurch Varianzen zu erkennen und auszugleichen, um somit aussagekräftige Messergebnisse zu erhalten. Basierend auf der theoretischen Höchstgeschwindigkeit von ca. 225.000 km/s, die Pakete in Netzkabeln erreichen können, ist die Zeit in Mikrosekunden zu messen, um den Fehler bei der Transformation auf Distanzen unter 225 Meter zu begrenzen. Die gesendeten Datenpakete folgen den Protokollen *Internet Control Message Protocol* (ICMP), *User Datagram Protocol* (UDP) und *Transmission Control Protocol* (TCP), indem Echo Requests über Ping und Traceroute durchgeführt werden. Durch Messungen mittels verschiedener Protokolle und Parameter lässt sich der Einfluss von lokalen Firewalls und anderen speziellen Netzkonfigurationen identifizieren, sodass sich solche Auswirkungen in die Standortbestimmung mit einbeziehen lassen.

Die Verzögerung auf dem Netzpfad besteht aus einem deterministischen und einem stochastischen Anteil. Die deterministische Verzögerung ergibt sich aufgrund der minimalen Verarbeitungszeit und der Warteschlangenverzögerung jeder Netzkomponente sowie der Übertragungszeit und Ausbreitungsgeschwindigkeit der Netzverbindungen.

Diese ist fester Bestandteil für jeden gegebenen Pfad in Abhängigkeit der Hardware- und Anschlussbedingungen sowie der Entfernung im Netz. Der stochastische Anteil in der Verzögerung entsteht durch Zwischenspeicherung und Warteschlangen bei der Abarbeitung paralleler Kommunikationsverbindungen sowie variabler Verarbeitungszeiten an jedem Zwischenrouter.

Die deterministische Verzögerung stellt die minimale Übertragungsverzögerung mit reduzierten Störeinflüssen dar. Für eine präzise Standortbestimmung sind stochastische Effekte zu vermeiden. Um diesen Effekt entgegenzuwirken, werden mehrere Messungen durchgeführt, um einen minimierten Wert nahe der deterministischen Latenz zu erhalten. Da sich der Einfluss stochastischer Effekte nicht ausschließen lässt, wird dies in den folgenden Berechnungen des Modell zur Geolokalisation berücksichtigt. Dazu erhält es eine weitere Komponente, welche die Wirkungen durch einen selbst-optimierenden Faktor berücksichtigt.

Abbildungsfunktion von Zeit und Entfernung

Da nicht die theoretische Signalgeschwindigkeit für die Ausbreitung der Daten in der Netztopologie angenommen werden kann, erfolgt eine Approximation durch Vergleichswerte. Dazu wird eine Abbildungsfunktion ermittelt, mit welcher die Umrechnung von der gemessenen Zeit in eine Entfernung erfolgt. Aus den Messergebnissen und den Arbeiten des *Centre for Applied Internet Data Analysis* (CAIDA) [287, 302] ist ersichtlich, dass die Korrelation zwischen Latenz und realer Distanz einer logarithmischen Kurve 5.1 folgt. Dies ist unter anderem auf die verschiedenen TIER Level in der Netztopologie zurückzuführen. Im Allgemeinen steht die physikalische Korrelation in einer direkten linearen Beziehung, sodass die Stärke der Kurvenkrümmung die Abweichung zwischen Modell und Realität darstellt, um die zuvor genannten stochastischen Einflüsse auszugleichen.

$$\mathit{Entfernung} = a \cdot \mathit{ln}(b \cdot \mathit{RTT} + c) + d \tag{5.1}$$

Die Parameter a, b, c und d der vollständig parametrisierten Funktion sind entsprechend der Trainingsdaten zu berechnen. Hierbei dient die Kommunikation zwischen den VPs als Referenz, wobei von diesen der Standort bekannt ist. Die Kurvenrekonstruktion erfolgt mit der minimierten Summe des quadratischen Fehlers. Ein Beispiel für eine solche individuelle Abbildungsfunktion ist in der Abbildung 5.18 dargestellt.

Es ist jedoch davon auszugehen, dass diese Werte von der VP-zu-VP Messung auf Grund der optimierten Position zu optimistisch sind. Um dem entgegenzuwirken wird das Ergebnis der Abbildungsfunktion bei der Zielbestimmung anschließend mit einem Faktor $0 < \lambda < 1$ multipliziert. In dem Zusammenhang bewirkt dies das Verschieben der logarithmischen Kurve auf der y-Achse nach unten. Der Faktor λ wird anhand des Abstandes von den Schnittpunkten der Multilateration automatisch angeglichen.

Die Werte zur Bestimmung der Abbildungsfunktion werden aus den RTTs und der Anzahl der Hops gemäß Gleichung 5.2 ermittelt. Die Verzögerung durch jeden einzelnen Hop auf dem Pfad wird aus der Latenz herausgerechnet, da in dieser Zeit das Signal keine geografische Entfernung zurücklegt. Dafür wird ein durchschnittlicher Wert

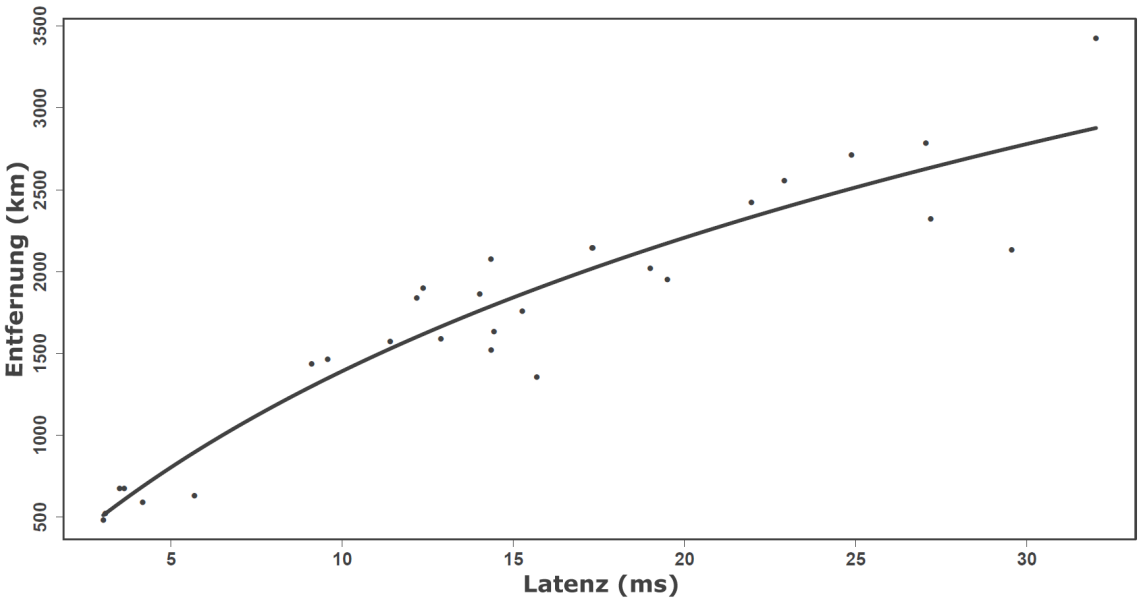


Abbildung 5.18: Angenäherte Abbildungsfunktion von Zeit und Entfernung mit der Summe der kleinsten Quadrate

von $T_{Hop} = 0,055\ ms$ angenommen [303], da dieser von vielen Faktoren wie der spezifischen Hardware und der Last auf dem Gerät abhängt. Zudem wird die initiale Verarbeitungszeit beim Senden und Empfangen des VP als auch des Ziels beim Erzeugen der Antwort herausgerechnet, indem dafür $2 \cdot T_{Hop} = 0,11\ ms$ von der RTT subtrahiert werden [29, 304].

$$Latenz_{Entfernung} = (RTT - 2 \cdot T_{Hop})/2 - T_{Hop} \cdot Hops \tag{5.2}$$

Die Krümmung der logarithmischen Kurve verdeutlicht die Übertragungsverzögerung durch die verarbeitenden Netzkomponenten im Vergleich zur Signalausbreitungsgeschwindigkeit im Leiter und glättet den stochastischen Anteil der Übertagungsverzögerung. Dieser Effekt ist insbesondere im Bereich der sogenannten „Letzten Meile“ zur Anbindung der Endnutzer auf Grund der COTS Hardware ausgeprägt.

Da die vollständige Netztopologie komplex und nicht öffentlich bekannt ist, müssen die realen Weglängen über verschiedene Ansätze im Modell der Geolokalisation approximiert werden. Entsprechend der praktischen Verlegung von Netzkabeln ist davon auszugehen, dass diese entlang von Straßen gebaut sind. Somit lässt sich das Straßennetz zur Berechnung der Verbindungslänge als Ersatz für notwendige Teilabschnitte verwenden. Hierfür erfolgt die Berechnung der Entfernungen über den Kartendienst von Google Maps [238]. Falls dieser Service keine realistischen Werte liefert, wird auf andere Verfahren wie orthodrome Entfernungsberechnung zurück gegriffen. Im Bezug zum generischen Modell besteht die Möglichkeit verschiedene Entfernungsfunktionen zur Berechnung heranzuziehen und deren Auswirkungen auf die Genauigkeit zu überprüfen. Darüber hinaus wird dadurch die Analyse verschiedener Approximationsvarianten auf einem einheitlichen Ansatz in Bezug zum Modell der Geolokalisation geschaffen. Im Vergleich zu bisherigen eingesetzten Verfahren ist der Verbindungsweg entlang der

realen Netzinfrastruktur länger als eine direkt verlaufende Gerade oder orthodrome Linie zwischen den Messpunkten. Eine Unterbewertung der Entfernung führt somit zu einer höheren Abweichung zwischen realen Standort und ermittelten Standort. Insbesondere wird der Längenunterschied beim Vergleich einer direkten Geraden mit den orthodromen Ansatz nach dem *World Geodetic System* Referenz Ellipsoid in der Abbildung 5.19 ersichtlich.

Abbildung 5.19: Gegenüberstellung der Entfernungsbrechnung von Frankfurt am Main nach Los Angeles gemäß einer direkten Verbindung und der orthodromen Linie (Vgl. [305])

Die Abbildungsfunktion von Zeit und Entfernung steht in Abhängigkeit zu den Verbindungen und der Konnektivität sowie den verwendeten Netzkomponenten und Netzmaterialien. Somit ist diese spezifisch für ein VP mit dessen Position im Netz. Die Funktion dient der Transformation von der gemessenen Latenz zur Ziel-IP-Adresse in eine Entfernung.

Lateration zur Standortbestimmung des Ziels

Um den unbekannten Standort einer IP-Adresse zu bestimmen, messen die ausgewählten VPs die Latenz und den Pfad zum Ziel mit unterschiedlichen Verfahren. Da die geographische Richtung bei der Kommunikation unbekannt ist, muss von einer Signalausbreitung in alle Richtungen ausgegangen werden. Ein Messergebnis stellt somit einen geometrischen Kreise dar, wobei die bekannte Position des VP den Mittelpunkt (X_{VP}, Y_{VP}) definiert und der zugehörige Radius R_{VP} des jeweiligen Kreises sich durch die Latenz ergibt. Der geographische Standort wird anschließend mittels der Methode Lateration bestimmt, indem die Schnittpunkte der Kreise berechnet werden. Dazu erfolgt vorher eine Umrechnung der Latenz und der Anzahl von Hops über die Abbildungsfunktionen des jeweiligen VP in eine Entfernung. Dadurch ist von jedem VP die Entfernung zum Ziel bekannt. Die Gleichung 5.3 beschreibt dafür einen Kreis. In der Abbildung 5.20 ist an einem Beispiel die Bestimmung des Standortes für ein Ziel dargestellt. Für jedes VP ist ein Kreis mit entsprechendem Radius eingezeichnet.

$$(X_{Ziel} - X_{VP})^2 + (Y_{Ziel} - Y_{VP})^2 = R_{VP} \tag{5.3}$$

Die Berechnung der Schnittpunkte von zwei Kreisen erfolgt gemäß der Gleichsetzungsmethode. Die Formeln sind im Anhang F.1 beschrieben. Zur Erhöhung der Präzision und zur Bestätigung des Ergebnisses wird eine mehrfache Lateration durchgeführt,

Abbildung 5.20: Visualisierung der Standortbestimmung über Lateration durch mehrere VPs

welche die verschiedenen Informationen der VPs nutzt. Da der Radius in der Einheit von km genutzt wird und der Mittelpunkt im geographischen Grad angegeben ist, erfolgt eine Transformation in einen gemeinsamen Koordinatenraum auf Basis der Maßeinheit von Grad. Entsprechend der Schnittpunktbestimmung von den Kreisen wird der zweidimensionale euklidische Raum zur Lateration verwendet, indem die Geokoordinaten direkt darauf abgebildet werden. Ein Grad entspricht 113,325 km am Äquator für Breiten- und Längengrad. Entlang des Längengrades reduziert sich dieser Wert vom Äquator zu den Polen, siehe Abbildung 5.21 und 5.22. Durch die Teilung des Kreisradius in km durch diesen Faktor ergibt sich die Entfernung vom VP zum Ziel approximiert in Grad.

Unterschiedliche Fälle der Schnittpunktberechnung

Bei der Lateration gibt es vier verschiedene Fälle, die bei der Schneidung von zwei Kreisen auftreten. Diese sind entsprechend dem Kontext der Geolokalisation zu unterscheiden und zu behandeln.

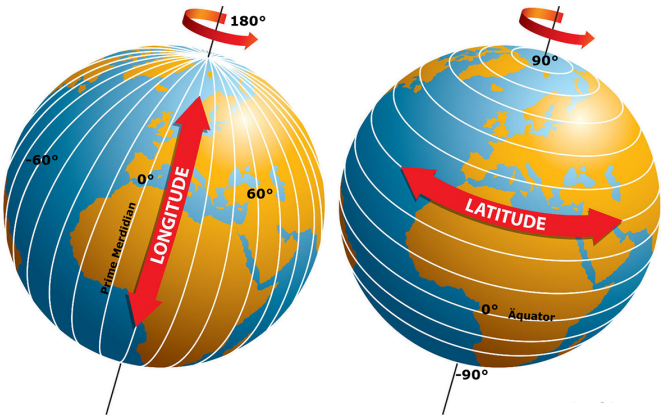


Abbildung 5.21: Unterschied zwischen Longitude und Latitude (Vgl. Dreamstime © Lukaves)

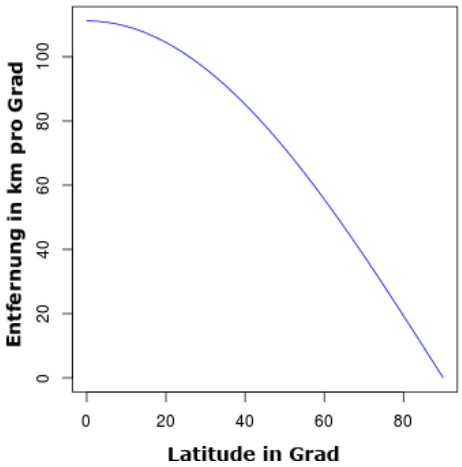


Abbildung 5.22: Verhältnis zwischen km und Grad entlang der Latitude (Vgl. [306])

1. Fall - Nicht überlappende Kreisflächen: Die Abbildung 5.23 zeigt die Situation, bei der sich die Kreisflächen nicht überlappen und es zu keinem Schnittpunkt kommt. Im Bezug zur Geolokalisation ist hierbei davon auszugehen, dass die ermittelten Radien zu kurz sind. Da von beiden VPs eine Messung zum Ziel erfolgte, ist davon auszugehen, dass sich der Standort des Ziels zwischen den zwei Kreisen befindet. Hierbei ist davon auszugehen, dass sich der Standort in der Mitte in Abhängigkeit der Größe von den Kreisradien befindet. Dazu werden die Radien von beiden Kreisen um den gleichen Faktor simultan vergrößert, sodass ein Schnittpunkt entsteht.

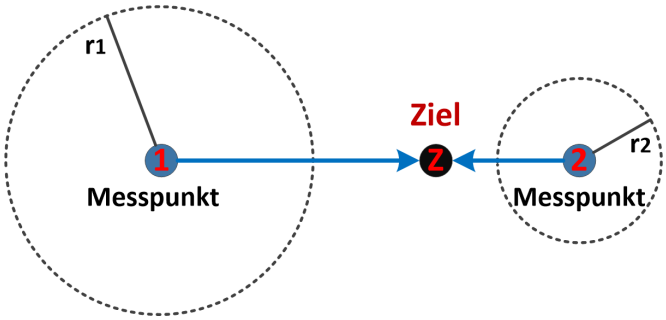


Abbildung 5.23: Kein Schnittpunkt bei nicht überlappenden Kreisen

2. Fall - Vollständig überlagernde Kreisflächen: Wenn ein Kreis vollständig von dem anderen überdeckt ist und somit innerhalb des zweiten Kreises liegt, dann gibt es ebenfalls keinen Schnittpunkt, siehe Abbildung 5.24. Hierbei ist im Sinn der Geolokalisation davon auszugehen, dass sich der Standort in der Region befindet, wo sich die Kreisradien am nächsten sind. Um einen Schnittpunkt zu erhalten, wird der Radius des größeren Kreises verkleinert, da das VP mit dem kleineren Radius näher am Ziel liegt und das genauere Ergebnis geliefert hat.

3. Fall - Angrenzende Kreisflächen: Beim idealen Fall berühren sich die Kreise in genau einem Punkt gemäß Abbildung 5.25. Dieser stellt den voraussichtlichen Standort des Ziels dar. In diesem Fall befinden sich die zwei VPs und das Ziel auf einer Geraden, wobei die Korrektheit in Abhängigkeit von potentiellen Störeinflüssen bei der Messung zu beurteilen ist.

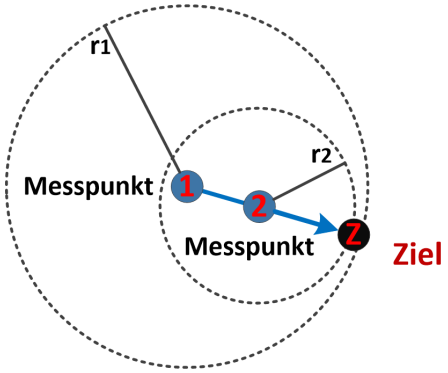


Abbildung 5.24: Kein Schnittpunkt bei überlagerten Kreisen

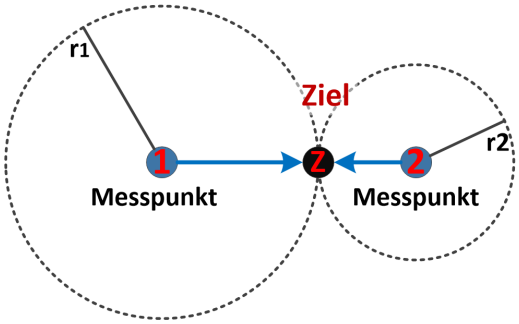


Abbildung 5.25: Ein Schnittpunkt bei angrenzenden Kreisen

4. Fall - Schneidende Kreisflächen: Wenn sich die VPs und das Ziel nicht auf einer Linie befinden, dann kommt es zu sich teilweise überlagernden Kreisen. Hierbei entstehen zwei Schnittpunkte, siehe Abbildung 5.26. Mit den derzeitigen Informationen lässt sich nicht entscheiden, an welcher dieser beiden Positionen sich das tatsächliche Ziel befindet. Daher werden beide Schnittpunkte als mögliche Standorte angenommen, sodass nachfolgend durch Filterung die Identifizierung des richtigen Standortes erfolgt. Bei der Auswertung aller Schnittpunkte ist zu erwarten, dass sich am korrekten Standort mehrere Schnittpunkte befinden und durch die dichte Gruppierung eine Bestätigung erfolgt. Dies ist in der Abbildung 5.27 dargestellt.

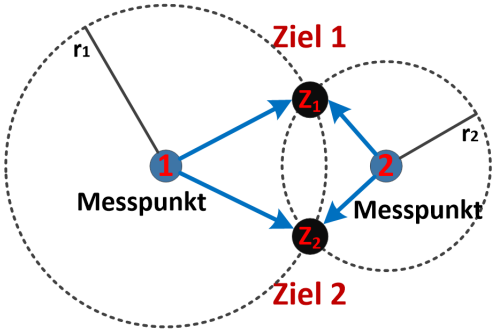


Abbildung 5.26: Zwei Schnittpunkte bei sich schneidenden Kreisen

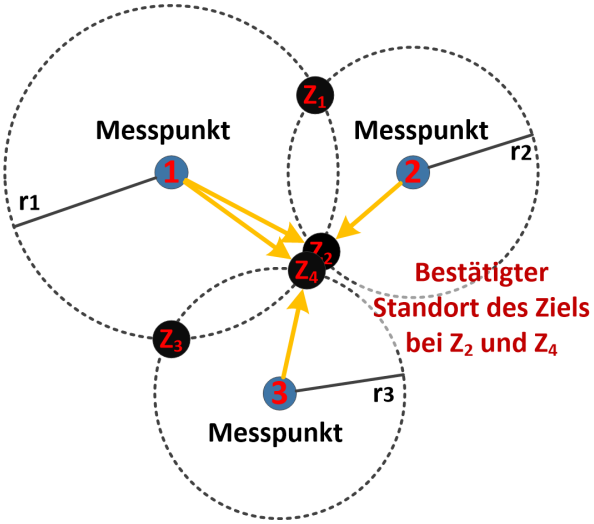


Abbildung 5.27: Lateration mit mehreren Kreisen und Schnittpunkten

Automatische Anpassung

Als Ergebnis der Multilateration ergibt sich eine Menge von Schnittpunkten, welche einer Art Punktwolke entspricht. Dies ist für ein reales Beispiel in der Abbildung 5.28 dargestellt. In Bezug zu den verschiedenen Fällen und zur Verfügung stehender Informationen erfolgt eine Priorisierung einzelner Punkte. Im 4. Fall ergeben sich zwei Resultate, wobei nur eins dem korrekten Wert entspricht. Zur gleichwertigen Berücksichtigung der Messwerte werden die entstandenen Schnittpunkte der Fälle 1 bis 3 doppelt gewichtet, da diese nur ein konkretes Ergebnis zurück liefern. Zur Modellierung und weiteren Verarbeitung dieser Punktewolke lässt sich das generische Modell verwenden, wobei die Werte mit dessen Standort durch Nutzer präsentiert sind. Dafür ist ein zentraler Standort zu finden, welcher der geografischen Position des Ziels entspricht.

Durch die Messung der VPs zu einem gemeinsamen Ziel entsteht die Mehrheit der Schnittpunkte an der selben Position beziehungsweise befinden sich auf Grund von Störeinflüssen dicht beieinander. Die Messabweichungen variieren auf Grund parallel stattfindender Kommunikation im Netz in gleichem Maße, sodass die Kreisradien der VPs im selben Ausmaß beeinflusst sind. Auf Grund von zusätzlichen Verzögerungen entstehen höhere Messwerte, was zu einer zu hohen Bewertung des Abstandes führt und letztendlich ungenaue Schnittpunkten ergibt. Um diesem Effekt entgegenzuwirken, erfolgt eine automatisierte Optimierung über den Faktor λ. Dazu wird von allen Schnittpunkten der Abstand zueinander mittels orthodromer Entfernungsfunktion ermittelt und beginnend nach dem kleinsten Wert sortiert. Auf Grund der maximalen Anzahl von zwei Schnittpunkten werden zur Identifizierung der relevanten Abstände alle Werte kleiner als der Median verwendet. Die Summe der Abstände gibt dabei Aufschluss, wie dicht die Schnittpunkte beieinander liegen. Je kleiner die Summe ist, desto dichter liegen die Punkte beieinander an nahezu denselben Standort und geben darüber die zu bestimmende Position des Ziels an. Zudem gibt die Summe auch Aufschluss über die Genauigkeit der Messergebnisse. Zur automatischen Anpassung wird ein iterativer Prozess durchgeführt, welcher die Multilateration mit reduzierten Kreisradien wiederholt und die Summe der Zwischenabstände berechnet. Die Reduktion der Kreisradien erfolgt über den Faktor λ mit Startwert $\lambda < 1$ entsprechend der Abbildung 5.29. Dies geschieht solange, bis sich die Summe nicht mehr verkleinert.

Schrittweise Filterung der korrelierten Informationen

Jeder ermittelte Schnittpunkt in der Menge stellt somit eine Wahrscheinlichkeit dar, dass sich an diesem Standort das gesuchte Ziel befindet. Hierbei besteht die Herausforderung in der Punktwolke den Bereich zu identifizieren, bei dem die meisten Punkte dicht beieinander liegen und die unbestätigten Scheinstandorte der Lateration herauszufiltern. Für den Bereich mit den dicht beieinanderliegenden Schnittpunkten ist anschließend der Mittelpunkt zu bestimmen, welcher dem geografischen Standort des Ziels entspricht. Über eine iterative Filterprozedur wird der dichte Bereich eingegrenzt. Dazu erfolgt die Bestimmung des Mittelpunktes hinsichtlich des Kriteriums der minimalen durchschnittlichen Entfernung über ein Clusterverfahren zur freien Platzierung. Als Metrik wird entsprechend der theoretischen Signallaufzeit die orthodrome Entfernungsfunktion

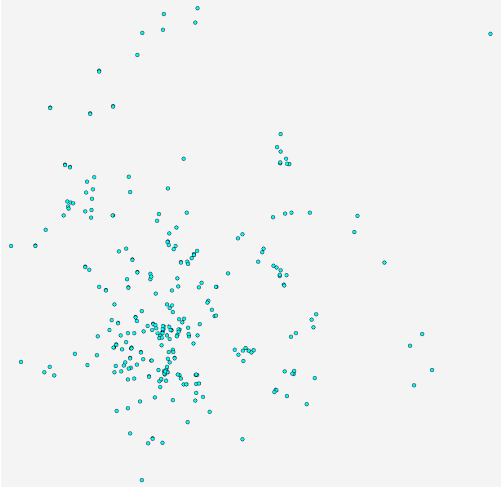


Abbildung 5.28: Punktwolke der Schnittpunkte mit einer Region dichter Standorte

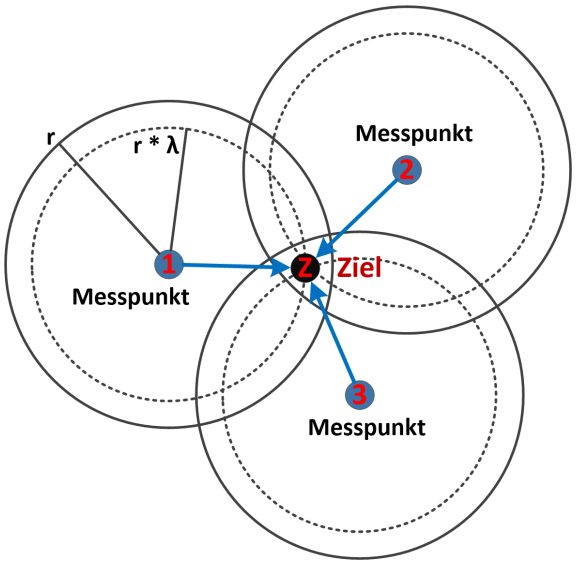


Abbildung 5.29: Automatische Anpassung der Kreisradien zum Ausgleich von Messungenauigkeiten

angewendet. Die Schnittpunkte mit der größten Entfernung zum Mittelpunkt werden in der nächsten Iteration nicht mehr berücksichtigt. Der eine Mittelpunkt nähert sich somit auf Grund des Kriteriums dem dichten Bereich an, sodass unbestätigte Standorte und Messergebnisse mit Störeinflüssen entfernt werden. Dieser Prozess der Filterung wird iterativ wiederholt, bis die Anzahl der verbleibenden Punkte kleiner gleich der Anzahl der verwendeten VPs liegt. Somit ergibt sich höchstens für jedes VP ein prognostizierter Standort des Ziels. Diese verbleibende Menge an Punkten repräsentiert den dichten Bereich, wobei der geografische Ort des Ziels in der Mitte zu erwarten ist. Der Abstand zwischen der ermittelten Lage und dem entferntesten Punkt im dichten Bereich wird als die erwartete Fehlerabweichung gesehen.

Spezifikation der Messung und Auswahl der Messergebnisse

Bei der Analyse mehrfacher Messungen sind maßgeblich die Messergebnisse mit der minimalen RTT bei reduzierten Störeinflüssen von Bedeutung, um den realen Abstand zu erhalten. In diesem Zusammenhang ist ein konstanter Netzpfad mit geringer RTT zur Messung geeignet. Neben der Latenz wird die Anzahl der Hops bestimmt. Dafür werden unterschiedliche ISO/OSI Layer 3+ Protokolle eingesetzt, um Netzknoten zu erkennen, da Router und Firewalls unterschiedlich auf ICMP, UDP und TCP Datenpakete reagieren [278]. Vor der Verwendung der Messergebnisse zur Geolokalisation, werden die Verbindungsdaten nach der minimalen Anzahl an Hops in Abhängigkeit des Protokolls gefiltert. Da jeder Hop sich mit einer stochastischen Verzögerung durch die Verarbeitung auf das Messergebnis auswirkt, erfolgt dadurch die Identifizierung des Netzpfades mit geringen Störeinflüssen. Für diese Verbindung wird von den verschiedenen Protokollen der maximale Wert für die Anzahl an Hops zur weiteren Verarbeitung verwendet, um möglichst der richtigen Anzahl von Netzknoten auf dem Pfad zu entsprechen. Nicht

jeder Netzknoten in einer Verbindung ist durch aktive Messung identifizierbar. Weitere Störeinflüsse durch asymmetrische Verbindungen und unterschiedliche Netzpfade werden durch mehrfache Messungen und die automatisierte Optimierung über den Faktor λ kompensiert.

Grenzen der aktiven Geolokalisation

Die Ermittlung von Standorten für die VPs erfolgt auf Grundlage der Infrastruktur. Hierbei werden verschiedene Medien und Netzkomponenten zur Übertragung eingesetzt, welche unterschiedliche Ausbreitungs- und Verarbeitungsgeschwindigkeiten aufweisen. Wenn diese nicht vollständig bekannt sind, dann lassen sich zentrale Knotenpunkte nicht optimal bestimmen. Dies beeinflusst zudem die notwendig Anzahl von VPs zur Messung. Zudem führen helix- und spiralförmige Infrastrukturen zu suboptimalen Berechnungen und dadurch zu höheren Abweichungen bei der Standortbestimmung. Weiterhin lassen sich im Zusammenhang mit Geolokalisation nur Ziele mit einem Standort innerhalb des minimal umgebenden Rechtecks der VPs genau bestimmen, da sonst eine Messung aus verschiedenen Richtungen nicht möglich ist.

5.3 Optimierungsalgorithmus Dragoon

Unabhängig des spezifischen Anwendungsgebietes erfolgt zur Platzierung und Identifizierung der zentralen Knotenpunkte die Entwicklung eines neuen Optimierungsalgorithmus Dragoon (Diversification Rectifies Advanced Greedy Overdetermined Optimization N-Dimensions). Passend zur Problemstellung bezeichnet das Akronym Dragoon historisch berittene Infanterie und ist abgeleitet vom Pferdenamen Dragoner. Diese Pferde wurden im 17. Jahrhundert primär für den schnellen Transport von den europäischen Ländern genutzt.

Der Optimierungsalgorithmus besteht aus einer Methode zur Initialisierung für eine gültige Anfangslösung und der Optimierung eines bestehenden Ergebnisses. Mit Wissen aus schrittweisen Analysen anderer Strategien erfolgt die Konstruktion eines modifizierten, hybriden Algorithmus nach dem Vorbild vom Zwei-Approx, Greedy und k-Means. Hierbei wird von einem Szenario ausgegangen, bei dem noch keine Zentren existieren. Die genannten Algorithmen lassen auch sich an Szenarien mit bereits fest vorgegebenen Standorten von Zentren jeweils anpassen.

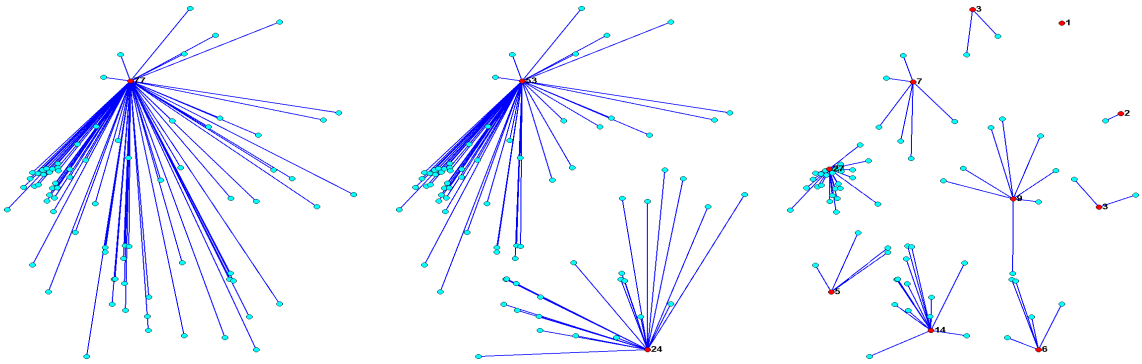
5.3.1 Analyse etablierter Strategien

Die analysierten Optimierungsstrategien bestehen aus zwei Phasen, der Initialisierung und der anschließenden Optimierung. Bei der Initialisierung wird eine valide Anfangslösung erzeugt. Hierbei sind die einzelnen Knoten dem nächstgelegenen Zentrum zuzuweisen und diese in Gruppen einzuteilen. Anschließend erfolgt eine iterative Optimierung, in dessen die initialisierten Positionen der Zentren verbessert werden.

Die Leistung der etablierten Algorithmen wird durch die initiale Platzierung der Zentren beeinflusst. Die Analyse der Greedy und der Zwei-Approx Strategie hat gezeigt, dass bei deren schrittweisen Initialisierung das zuerst platzierte Zentrum eine hohe Anzahl von Knotenpunkten nach Abschluss der Initialisierung bindet im Vergleich zu den restlichen platzierten Zentren. Insbesondere die wiederholte Ausführung des Zwei-Approx Algorithmus mit der zufälligen Platzierung des ersten Zentrums weist dieses Verhalten auf. Das ist in der Planungsphase von den Zentren unproblematisch, speziell bei theoretischen Systemen mit unbeschränkter Bedienkapazität. Dennoch ist eine gleichmäßige Verteilung von Knotenpunkten zu den Zentren auf Grund einer gleichmäßigen Lastverteilung erstrebenswert. Diese Ansätze zur Initialisierung zeigen den Einfluss der ersten Platzierungsentscheidung auf die Gesamtlösung, wodurch auch die endgültige Lösung maßgeblich beeinflusst wird. In der Untersuchung der k-Means Strategien [182, 184, 185] mit der gleichzeitigen, zufälligen Initialisierung der Zentren wurde festgestellt, dass eine knotenbasierte Platzierung für die anschließenden Optimierungsschritte vorteilhaft ist. Bei einer zufälligen Platzierung kommt es vor, dass einzelne Zentren zu dicht beieinander liegen. Die Zentren müssen daraufhin zunächst auseinander gesetzt werden, um eine Verteilung mit lokalem Optima zu finden und optimierte Gruppen zuzuordnen. Bei einer freien Platzierung während der Initialisierung werden die Standorte nicht berücksichtigt. Insbesondere beim Llyod Algorithmus kann es vorkommen, dass ein Zentrum außerhalb der effektiven Reichweite liegt und keine Auswirkung hat. In der Optimierungsphase sind dadurch mehr Optimierungsschritte notwendig als bei einer gleichmäßigeren Verteilung der initialisierten Zentren. Dieses Verhalten wirkt sich auf die benötigte Rechenzeit aus. Der Aspekt wird insbesondere beim k-Means++ berücksichtigt, indem die Wahrscheinlichkeit solcher Konstellationen verringert wird. Um den Einfluss der ersten Platzierungsentscheidung zu vermindern sowie die Sensitivität von Zufall zu reduzieren, wird bei Dragoon die initiale Lösung mittels einer neuen und vorwiegend deterministischen Methode nach dem Ansatz von Zwei-Approx definiert.

Während der Optimierungsphase versuchen die k-Means Algorithmen nur eine optimierte Lösung in Bezug auf eine einzelne Gruppe von Knoten mit nur einem Zentrum zu erreichen. Dadurch geht der gegenseitige Einfluss zu anderen Gruppen und zum Gesamtsystem verloren, sodass dieser Ansatz von vornherein auf ein lokales Optima und suboptimale Lösungen abzielt.

Die Abbildungen 5.30, 5.31, 5.32 und 5.33 zeigen den Unterschied während der Platzierung am Beispiel des Szenarios Deutschland und die damit verbundenen Probleme. Mit dem Wissen dieser Zusammenhänge und Defizite gegenwärtiger Optimierungsansätze wird ein neues Optimierungsverfahren designt, welches auch für unterschiedliche Nebenbedingungen geeignet ist. Darüber hinaus wird der Einfluss von zufälligen Entscheidungen vermindert, sodass deterministische Ergebnisse entstehen und mehrfache Ausführungen vermieden werden.

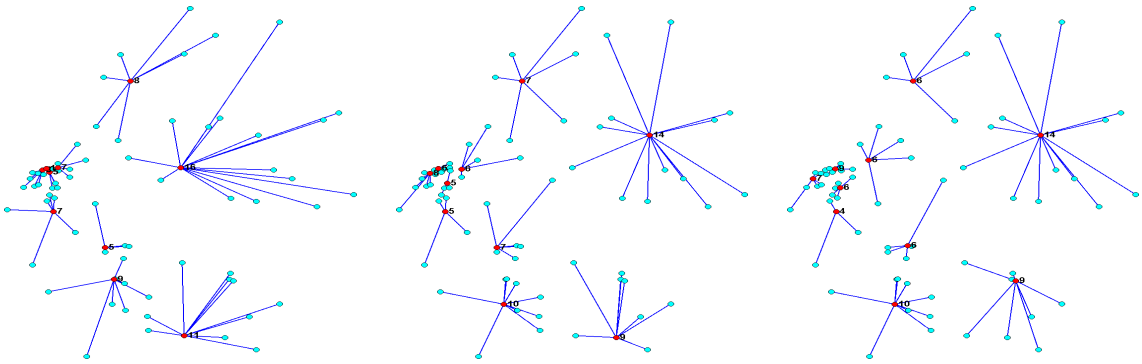


(a) 1. Zentrum an zufälliger Position

(b) 2. Zentrum an der am weitest entfernten Position

(c) Platzierung am Ende der Optimierung

Abbildung 5.30: TwoApprox mit der Platzierung von zehn Zentren

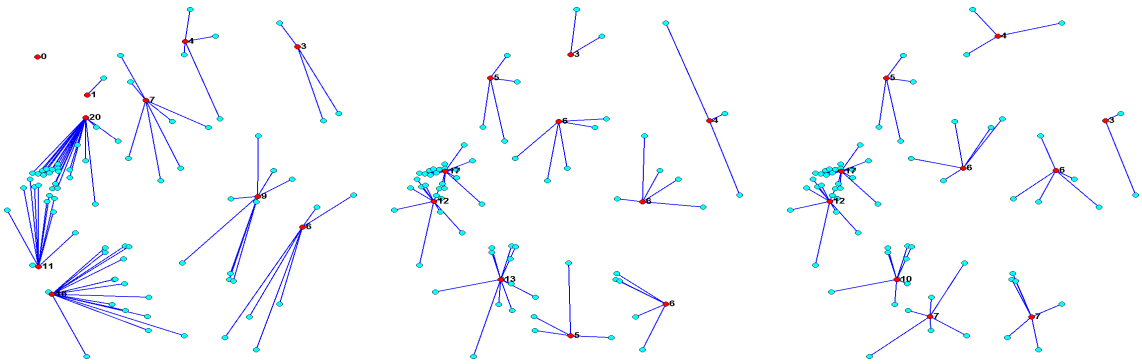


(a) Zufällige Platzierung von zehn Zentren am Knoten

(b) Lokale Optimierung der Standorte von den Zentren und erneute Zuweisung

(c) Platzierung am Ende der Optimierung

Abbildung 5.31: MacQueen mit der Platzierung von zehn Zentren

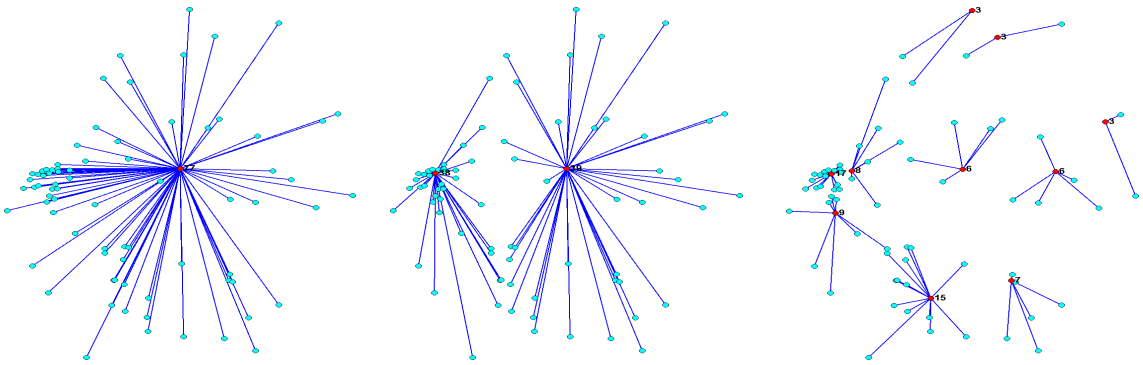


(a) Zufällige Platzierung von zehn Zentren an beliebiger Position

(b) Lokale Optimierung der Standorte von den Zentren und erneute Zuweisung

(c) Platzierung am Ende der Optimierung

Abbildung 5.32: Lloyd mit der Platzierung von zehn Zentren



(a) Zielgerichtete Platzierung des ersten Zentrums

(b) Platzierung des zweiten Zentrums an der nächstbesten Position

(c) Platzierung am Ende der Optimierung

Abbildung 5.33: Greedy mit der Platzierung von zehn Zentren

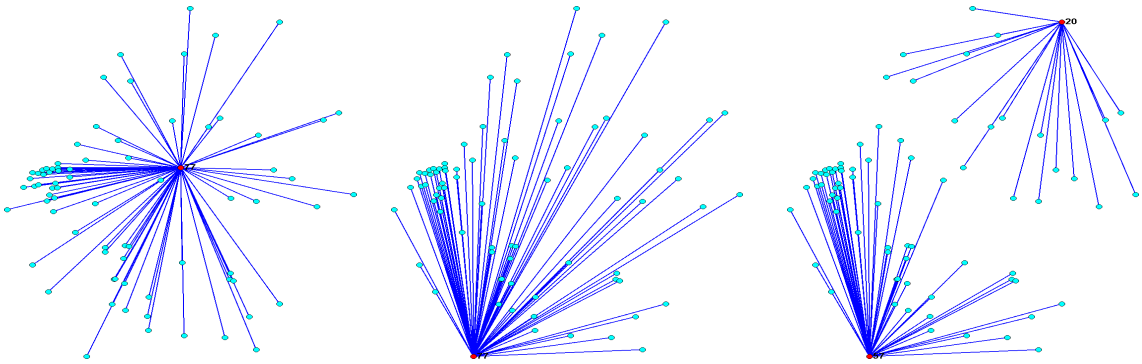
5.3.2 Initialisierung

Basierend auf den Analysen der bekannten Strategien zur Initialisierung wurde eine neue Methode entwickelt, welche auf der Zwei-Approx Strategie beruht und geringen Zufallseinfluss hat. Während der Initialisierung erfolgt die Platzierung der Zentren stets nur mit der Nebenbedingung Knotenplatzierung. Zu Beginn der Initialisierungsphase wird ein sogenannter Orientierungspunkt bzw. ein Scheinzentrum entsprechend dem 1-Zentrums Problem platziert. Hierbei gibt es nur ein Zentrum für alle Knotenpunkte, welches an der optimalen Position in Relation zu allen Knoten ist. Dafür ist eine einmalige vollständige Durchsuchung des Lösungsraums notwendig, wobei es sich in diesem Fall mit bidirektionalen Verbindungen um eine quadratische Komplexität $\mathcal{O}(v \cdot (v-1))$ handelt. Dies ist vergleichbar mit der Platzierung des ersten Zentrums beim Greedy bzw. Brute Force Algorithmus. Da diese Berechnung nur einmalig auszuführen ist, hält sich der Berechnungsaufwand in Grenzen. Der Orientierungspunkt ist nicht Teil der finalen Lösung.

Anschließend erfolgt die eigentliche Platzierung der vorgegebenen Anzahl von Zentren, wobei der Orientierungspunkt zur Anzahl nicht dazu zählt. Die iterative Platzierung erfolgt nach dem adaptierten Prinzip der Zwei-Approx Strategie [54]. Dazu wird für jeden Knoten die Entfernung zu allen Zentren des Szenarios berechnet. Im ersten Schritt ist noch kein Zentrum vorhanden, sodass die Entfernungsberechnung zum Orientierungsknoten erfolgt. Dies steht entgegen der zufälligen Wahl des ersten Zentrums bei der Zwei-Approx Strategie. Der Orientierungspunkt wird nach der Platzierung des ersten Zentrums wieder entfernt, sodass bei allen weiteren Platzierungsentscheidungen der Orientierungspunkt keine Beachtung mehr findet. Das nächste Zentrum wird stets an dem selben Standort des Knotens platziert, welcher die größte Entfernung zu seinem nächstgelegenen Zentrum aufweist. Falls mehrere Standorte mit dieser Eigenschaft vorhanden sind, wird einer davon zufällig ausgewählt. Dieser Vorgang wiederholt sich solange, bis die vorgegebene Anzahl an Zentren gesetzt ist. Nach jeder Platzierung eines neuen Zentrums, erfolgt eine Überprüfung der Entfernungen zum nächstgelegenen Zentrum. Dafür sind $\sum_{i=1}^{k}(i \cdot v)$ Berechnungen notwendig. Diese entsprechen der Komplexität $\mathcal{O}(k \cdot v)$.

Durch das Entfernen des Orientierungsknotens wird eine Initialisierungslösung generiert, bei der eine ausgeglichenere Zuweisung von Nutzern zu den einzelnen Zentren entsteht, als bei den Verfahren MacQueen und Lloyd. Weiterhin wird durch diese Initialisierung am Knoten erreicht, dass sich die Zentren bereits ab der ersten Standortentscheidung am Rand des Platzierungsraumes befinden und die Zentren möglichst weit voneinander entfernt sind. Dies führt zu einer gleichmäßigen Verteilung der Zentren im Platzierungsraum.

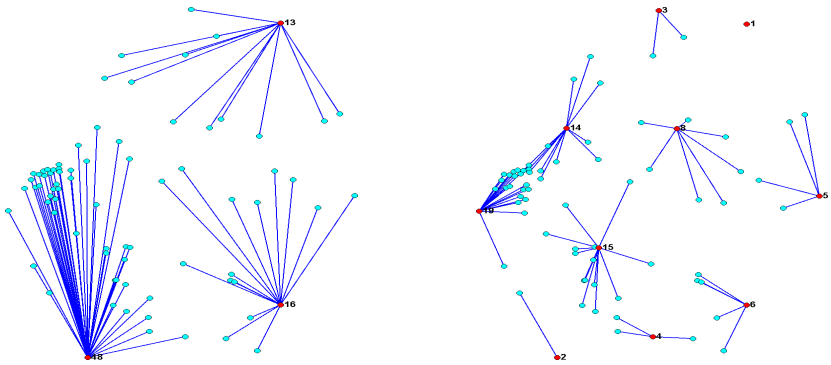
Die Zentren liegen im Bezug zur zugeordneten Knotengruppe am Rand dieser Gruppe bzw. an einer für die Gruppe nicht optimalen Position. Dies bietet weiteres Optimierungspotential nach Abschluss der Initialisierung, welches in der anschließenden Optimierungsphase genutzt wird. Die Abbildung 5.34 zeigt die Schritte der Initialisierung am Beispiel des Szenarios von Deutschland. Durch diesen Ansatz entsteht eine spezifische Lösung des Zwei-Approx Algorithmus. Somit bleibt auch die zwei-approximierte Garantie der Initialisierungslösung gültig [78, 176]. Nach der Initialisierung wird die berechnete Zwischenlösung mittels einer iterativen Optimierung weiter verbessert.



(a) Zielgerichtete Platzierung des Orientierungspunktes

(b) Platzierung des ersten Zentrums an der am weitesten entfernten Position

(c) 2. Zentrum an der am weitesten entfernten Position



(d) Platzierung aller weiteren Zentren an der stets am weitesten entfernten Position

(e) Platzierung am Ende der Initialisierung

Abbildung 5.34: Dragoon mit der Initialisierung von zehn Zentren

5.3.3 Optimierung

Für die schrittweise Optimierung wird ein veränderter und verallgemeinerter k-Means Algorithmus mit einer umfassenderen Sichtweise auf die Problemstellung eingesetzt. Der folgende Abschnitt beschreibt den allgemeinen Optimierungsansatz, welcher anschließend an die Nebenbedingungen der Standortwahl aus Kapitel 2.3.2 adaptiert wird. Die Optimierung der Standorte erfolgt iterativ, wobei eine Iteration in einzelne Schritte für ein jeweiliges betrachtetes Zentrum unterteilt ist. Dabei wird für jedes Zentrum anhand aller Nutzer ein aktualisierter Standort berechnet.

Zu Beginn einer Iteration wird festgelegt, in welcher Reihenfolge die Betrachtung der Zentren erfolgt. Dazu werden die Zentren entsprechend nach der längsten Entfernung der jeweils zugewiesenen Nutzer sortiert, beginnend mit dem Zentrum mit der größten Distanz zu einem Nutzer. Dafür sind $\leq \mathcal{O}(k \cdot log\ k)$ Rechenoperationen notwendig. In einem Optimierungsschritt testet der Algorithmus für ein Zentrum neue Standorte um die gegenwärtige Position. Falls durch eine Nebenbedingung eine Infrastruktur durch einen Graphen vorgegeben ist, geschieht dies in Abhängigkeit der angrenzenden Verbindungen. Anderenfalls werden die getesteten Möglichkeiten in dem Schritt auf den gegenwärtigen Bereich begrenzt, um die Anzahl der Möglichkeiten einzugrenzen und die Optimierungszeit zu verringern. In Abhängigkeit der durchschnittlichen Anzahl $dat = \frac{v}{k}$ getesteter Positionen für ein Zentrum sind $\mathcal{O}(dat \cdot v)$ Auswertungen notwendig. Während der Prozedur erfolgt bei jedem Testschritt eine erneute Zuweisung aller Knoten im Szenario zu deren nächstgelegenen Zentren, da sich eine Änderung der Platzierung entsprechend auf die Knotenzuweisungen auswirkt. Dazu wird überprüft, ob sich die Entfernung zu dem neuen Zentrum gegenüber dem gegenwärtigen verbessert. Dies führt dazu, dass stets der aktuelle Gesamtzustand ersichtlich ist.

Die Position mit dem besten Zielfunktionswert wird als neuer Standort des betrachteten Zentrums verwendet. Die Bewertung geschieht immer in Relation zum Gesamtszenario, da eine lokale Änderung auf Grund des Zusammenhangs einen globalen Einfluss auf das gesamte System hat. Dies ist auch durch die NP-schwere des Problems bedingt.

In jedem Optimierungsschritt verwendet der Algorithmus stets den aktuellen Standort eines jeden Zentrums zur Bewertung einer neuen Platzierungsentscheidung für ein anderes Zentrum. Die Auswertung des Szenarios erfolgt gemäß dem spezifizierten Optimierungskriterium. Falls bei der Evaluierung eines Szenarios keine Veränderung des primären Optimierungskriteriums ersichtlich ist, verwendet der Algorithmus ein sekundäres Optimierungskriterium. Dieses dient dem Erkennen von kleineren Veränderungen, sodass sich zwischen zwei Lösungen mit gleichem primären Fitnesswert eine genauere Unterscheidung treffen lässt. Dazu dient der Durchschnittswert aus allen Zielfunktionswerten von den Zentren. In jeder Iteration lässt sich die Position eines Zentrums nur einmal verändern, sodass die Zentren untereinander gleichberechtigt in deren Standortänderungen sind. Dies führt, wie auch die Fokussierung auf angrenzende Standorte um die gegenwärtige Position, zu einer schrittweisen Akzeptanz von Verbesserungen und vermeidet eine zu schnelle Stagnation in einem lokalen Optimum.

Die iterative Optimierung wird solange wiederholt, bis sich alle Standorte der Zentren in einer Iteration nicht mehr ändern. Auf Grundlage der vorausgegangenen Initialisierung

sind weniger Iterationen als bei den anderen k-Means Verfahren notwendig, bis der Algorithmus terminiert. Da während der Optimierungsphase nur verbesserte Positionen akzeptiert werden, bleibt die zwei-approximierende Güte erhalten und der Algorithmus wird immer terminieren. In Abhängigkeit der Anzahl durchzuführenden Iterationen *iter* ergibt im Durchschnitt insgesamt für die Optimierung eine Komplexität von $\mathcal{O}(iter \cdot v^2)$ und somit eine quadratische Laufzeit für den Algorithmus Dragoon.

5.3.4 Adaption an Nebenbedingungen

Die vorausgegangene Meta-Beschreibung dient dem allgemeinen Verständnis über den Ablauf des Algorithmus. In Abhängigkeit der Infrastruktur und der Nebenbedingungen zur Platzierung von Zentren sind die Schritte der Optimierung an die Gegebenheiten anzupassen. Durch die spezifische Adaption lässt sich das Potential der Platzierungsmöglichkeiten im vollem Umfang ausnutzen, sodass der heuristische Ansatz bestmögliche Ergebnisse erreicht. Die Methode zur Ermittlung einer Initialisierungslösung ist für die verschiedenen Nebenbedingungen zur Platzierung von Zentren stets gleich. Da der Initialisierungsprozess die Zentren am Knoten platziert, genügt dies den restriktivsten Anforderungen gemäß der knotenbasierten Platzierung und ist somit immer anwendbar. Die folgende Differenzierung in der Optimierungsphase spezifiziert eine Platzierung gemäß:

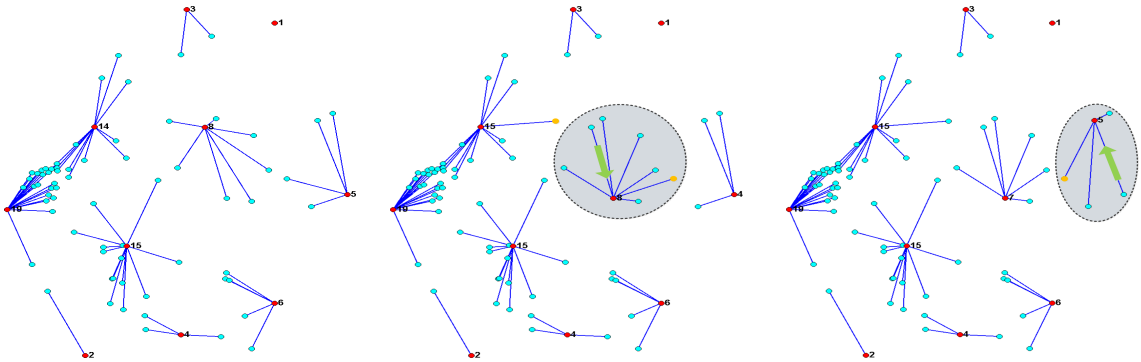
- Knotenbasierte Platzierung
- Infrastrukturbasierte Platzierung
- Freie Platzierung

Knotenbasierte Platzierung

Bei der knotenbasierten Platzierung gelten nur die vordefinierten Standorte als valide Lösungen zur Platzierung eines Zentrums. Während der Optimierungsphase werden daher auch nur valide Zwischenlösungen akzeptiert. Hierbei ist dem Algorithmus keine Infrastruktur explizit bekannt und die Entfernungsberechnung erfolgt transparent. Nach der Initialisierung sind die Knoten des Szenarios den verschiedenen Zentren zugewiesen und somit in Gruppen eingeteilt. Die Standorte der jeweils einem Zentrum zugewiesenen Knoten werden in einem Optimierungsschritt als neuer Standort geprüft. Wenn der neue Standort für das betrachtete Zentrum zu einer Verbesserung der Gesamtsituation führt, wird dieser übernommen. Anderenfalls bleibt das Zentrum am vorherigen Standort bestehen. Zur Identifikation eines besseren Standortes wird das gesamte Szenario evaluiert. Dies beinhaltet die wiederholte Zuordnung aller Knoten zu dem jeweiligen Zentrum mit der geringsten Entfernung. Dazu wird von allen Zentren der zuletzt ermittelte Standort verwendet, ausgenommen von dem betrachteten Zentrum, welches mit einem neuen Standort getestet wird.

Bei der knotenbasierten Platzierung sind die Möglichkeiten für einen besseren Standort in jeder Iteration auf dessen zugeordnete Gruppe limitiert. Der Berechnungsaufwand in einer Iteration ist linear abhängig von der Anzahl der getesteten Standorte. Zur weiteren

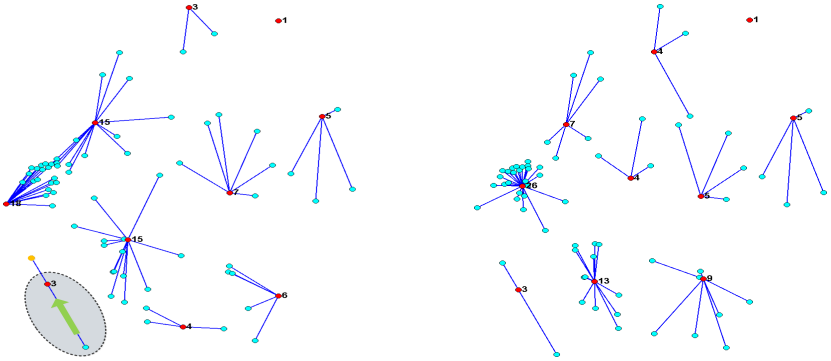
Reduzierung des Aufwandes lassen sich die zu testenden Knoten einer Gruppe auf die zum Zentrum nächstgelegenen eingrenzen. Der Algorithmus optimiert die Positionen der Zentren schrittweise bis es keine Veränderung mehr in einer Iteration gibt. Angelehnt an die k-Means Strategien und dem Teile-und-Herrsche Prinzip wird die Lösung des gesamten Szenarios durch die Berechnung der Teillösungen erreicht. Im Gegensatz zu den bisherigen Ansätzen von k-Means bleibt durch die umfassende Evaluierung der Gesamtsystembezug erhalten. Die Abbildungen 5.35 und 5.36 zeigen anhand eines Beispiels die knotenbasierte Platzierung von Zentren. Die markierten Bereiche verdeutlichen die Veränderungen.



(a) Platzierung am Ende der Initialisierung

(b) Optimierung des erstens Zentrums in der ersten Iteration

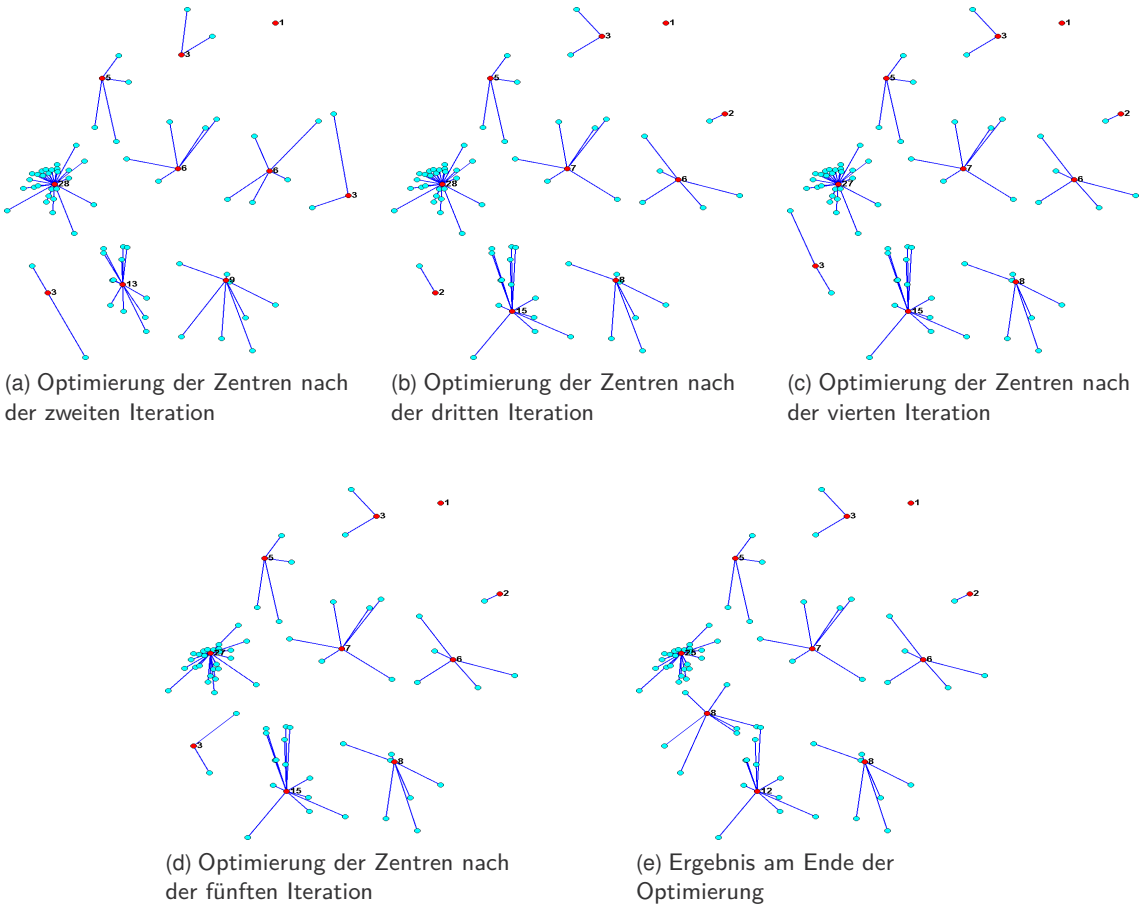
(c) Optimierung des zweiten Zentrums in der ersten Iteration



(d) Optimierung des dritten Zentrums in der ersten Iteration

(e) Optimierung aller Zentren nach der ersten Iteration

Abbildung 5.35: Erste Iteration der Optimierung von Dragoon mit der Nebenbedingung knotenbasierte Platzierung



(a) Optimierung der Zentren nach der zweiten Iteration

(b) Optimierung der Zentren nach der dritten Iteration

(c) Optimierung der Zentren nach der vierten Iteration

(d) Optimierung der Zentren nach der fünften Iteration

(e) Ergebnis am Ende der Optimierung

Abbildung 5.36: Optimierung mit Dragoon mit der Nebenbedingung knotenbasierte Platzierung

Infrastrukturbasierte Platzierung

Bei der infrastrukturbasierten Platzierung ist mit dem Szenario eine Verbindungstopology vorgegeben. Diese explizite Kenntnis über die Infrastruktur bezieht der Algorithmus mit in die Optimierung ein. Die Platzierung der Zentren erfolgt dabei an den Kreuzungspunkten der Verbindungen. Nach der Initialisierung sind Zentren an der Infrastruktur platziert, da jeder Knoten über diese miteinander verbunden ist. Gemäß dem beschriebenen, allgemeinen Optimierungsansatz probiert der Algorithmus in einem Iterationsschritt Standorte um die gegenwärtige Position. Bei der infrastrukturbasierten Platzierung probiert der Algorithmus in einem Optimierungsschritt für ein Zentrum die jeweils direkt über die Infrastruktur angrenzenden Kreuzungspunkte. Dadurch beschränkt sich die zu testende Anzahl auf eine geringe Menge, wodurch der Berechnungsaufwand begrenzt wird. Zur Auswertung der Lösung erfolgt wieder die Zuweisung aller Nutzer zu deren nächstgelegenen Zentrum. Der Standort mit der größten globalen Verbesserung wird als neuer Standort akzeptiert und ersetzt damit den Standort des gegenwärtig betrachteten Zentrums. Anderenfalls bleibt der Standort des Zentrums unverändert. Die Entfernungsberechnung erfolgt stets in Abhängigkeit der Infrastruktur und dessen Metrik. Durch die Abbildung auf eine Graphenstruktur lässt sich der kürzeste Verbindungsweg mittels Dijkstra's Algorithmus berechnen und über die Knotenanzahl bestimmen, wobei in Abhängigkeit der Verbindungseigenschaften eine Wichtung mit einzubeziehen ist. Die

Abbildung 5.37 verdeutlicht die Vorgehensweise vom Optimierungsalgorithmus bei der Nebenbedingung der infrastukturbasierten Platzierung. Für das Szenario Colt zeigen die Abbildungen 5.38 und 5.39 die Optimierung gemäß der infrastrukturbasierten Platzierung. Als Entfernungsmetrik wird die Anzahl der Hops verwendet, wobei bei gleicher Anzahl zufällig entschieden wird.

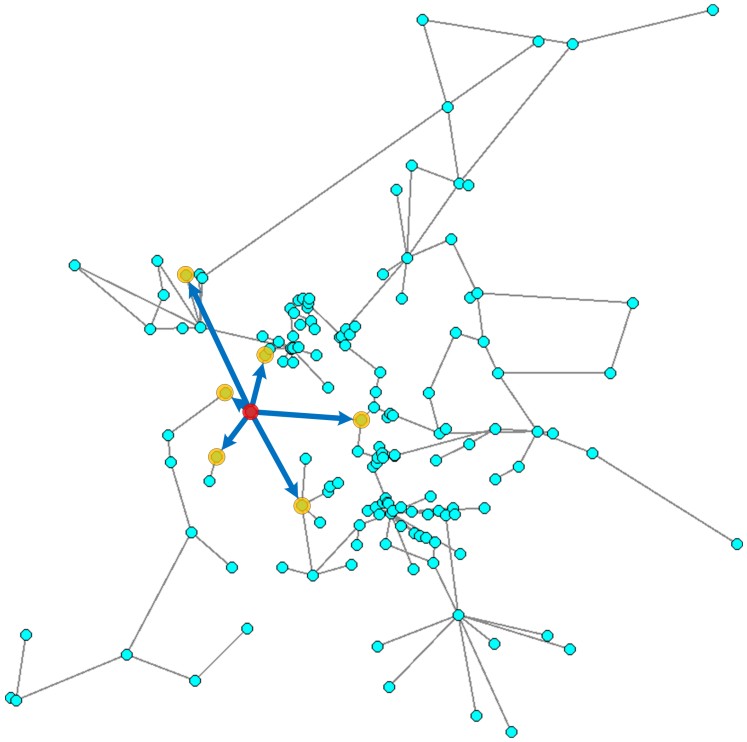
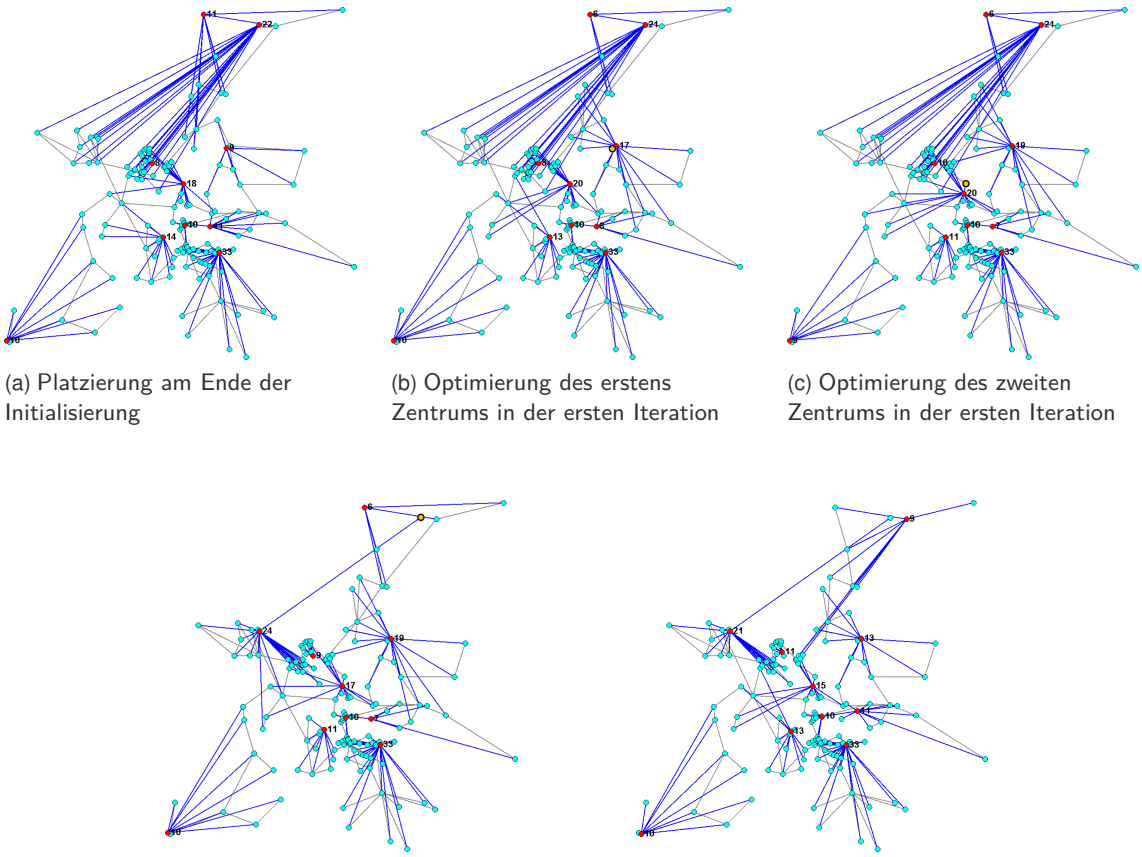


Abbildung 5.37: Verschiebung des aktuellen Standortes (rot) zu den Möglichkeiten der direkt benachbarten Knoten (gelb)



(a) Platzierung am Ende der Initialisierung

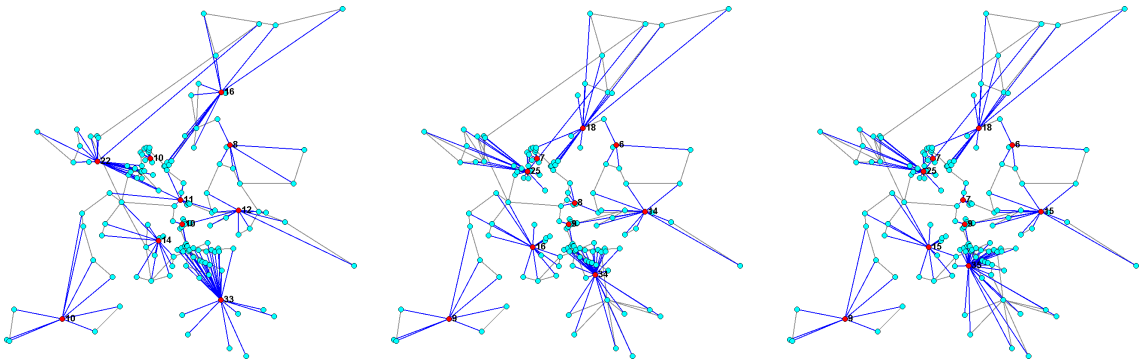
(b) Optimierung des erstens Zentrums in der ersten Iteration

(c) Optimierung des zweiten Zentrums in der ersten Iteration

(d) Optimierung des dritten Zentrums in der ersten Iteration

(e) Optimierung aller Zentren nach der ersten Iteration

Abbildung 5.38: Erste Iteration der Optimierung von Dragoon mit der Nebenbedingung infrastrukturbasierte Platzierung



(a) Optimierung der Zentren nach der zweiten Iteration

(b) Optimierung der Zentren nach der dritten Iteration

(c) Ergebnis am Ende der Optimierung

Abbildung 5.39: Optimierung mit Dragoon mit der Nebenbedingung infrastrukturbasierte Platzierung

Freie Platzierung

Bei der freien Platzierung lassen sich die Standorte der Zentren innerhalb des Platzierungsraumes frei festlegen. Somit ergibt sich eine unendliche Vielfalt an Möglichkeiten. Zur Handhabung und Eingrenzung wird der Platzierungsraum durch ein Gitternetz diskretisiert und damit ergeben sich wie bei der knotenbasierten Platzierung vordefinierte Standorte. Der Algorithmus prüft systematisch alle umliegenden Standorte mit einem spezifizierten Abstand ε um die gegenwärtige Position herum als neuen Standort. Dies erfolgt zunächst mit Begrenzung auf die dem Zentrum zugewiesene Gruppe. Anschließend überprüft der Algorithmus, ob dieser Standort einen besseren Zielwert für das gesamte Szenario erreicht und speichert diesen entsprechend ab. Dieser Vorgang wird solange wiederholt bis keiner der getesteten Standorte zu einer Verbesserung führt. Anschließend wird die Granularität des Gitternetzes sukzessiv verkleinert gemäß $\varepsilon_{neu} := \varepsilon_{alt} \ / \ 2$ und wieder alle umliegenden Standorte überprüft. Dieser Prozess wird wiederholt ausgeführt, bis eine vordefinierte Granularität erreicht wurde. Dazu überprüft der Algorithmus ob der Gitternetzabstand ε kleiner als die maximal akzeptierte Abweichung ist. Die definierte Schranke dient der Terminierung des Verkleinerungsprozesses von ε im Iterationsschritt und damit der Verschiebung eines Zentrums. Da der Wert von ε vom jeweiligen Szenario abhängig ist, wird dieser mittels einer Meta-Optimierung berechnet. Der Startwert für ε ergibt sich in Abhängigkeit der Anzahl der Zentren k und der größtmöglichen Entfernung innerhalb des Hüllkörpers vom Szenario, siehe Formel 5.4. Dieser Wert entspricht dem maximalen Radius eines Zentrums bei gleichmäßiger Verteilung im Platzierungsraum. Damit das Zentrum bei der ersten Verschiebung nicht an den gegenüberliegenden Rand des Hüllkörpers verschoben wird, ist dieser Wert noch zu halbieren. Die untere Schranke für die Granularität ist entsprechend der notwendigen Genauigkeit der Platzierung festzulegen.

$$\varepsilon = \frac{\sqrt{x^2 + y^2 + ...}}{k} \tag{5.4}$$

Die beiden Iterationsschritte Verschiebung eines Zentrums und Verkleinerung des Gitternetzes sind in der Abbildung 5.40 dargestellt. Die linke Seite zeigt die Verschiebung der aktuellen Position des Zentrums zu einem besseren Standort. Die rechte Seite verdeutlicht den Prozess der Verfeinerung des Gitternetzes und damit die genauere Positionierung.

Die Abbildungen 5.41 und 5.42 zeigen anhand eines Beispiels die freie Platzierung von Zentren. Die markierten Bereiche verdeutlichen die Veränderungen. Durch die verbesserte Positionierung von den Zentren ist es möglich, dass ein oder mehrere Zentren nach einer Iteration keine Nutzer mehr zugewiesen bekommt. In dieser Ausnahmesituation erfolgt eine Reinitialisierung des entsprechenden Zentrums an den Standort des Nutzers mit der größten Entfernung zu einem Zentrum.

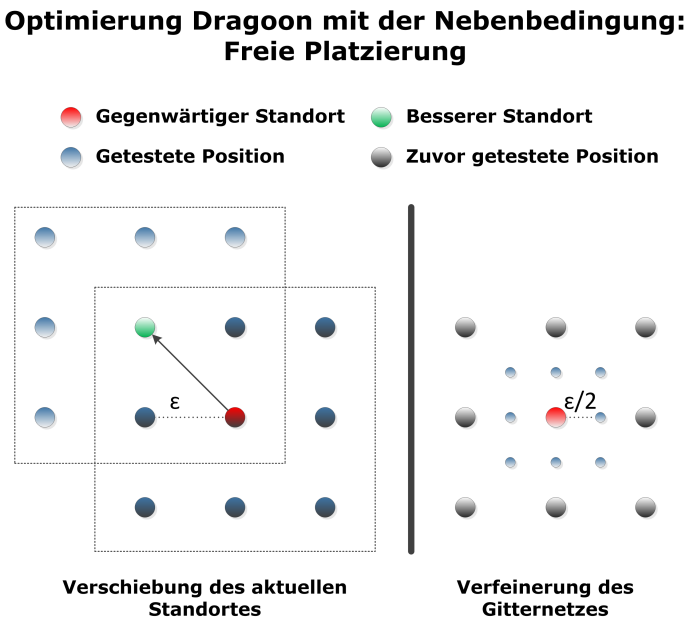
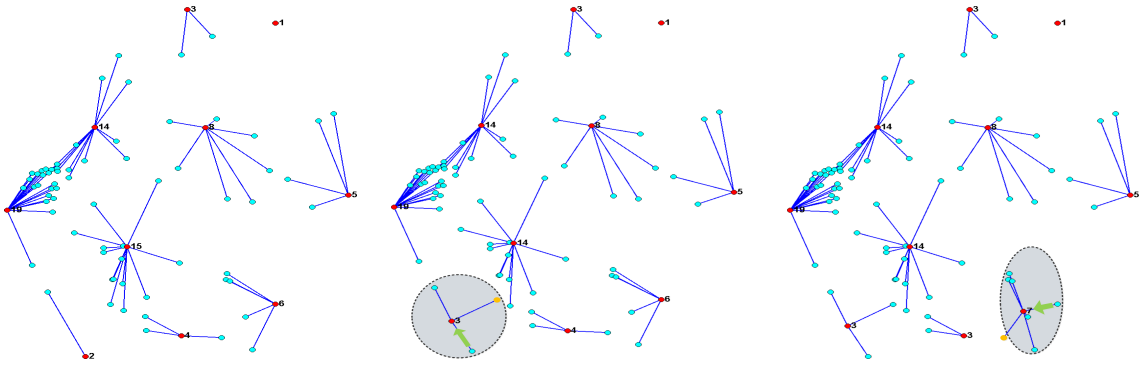


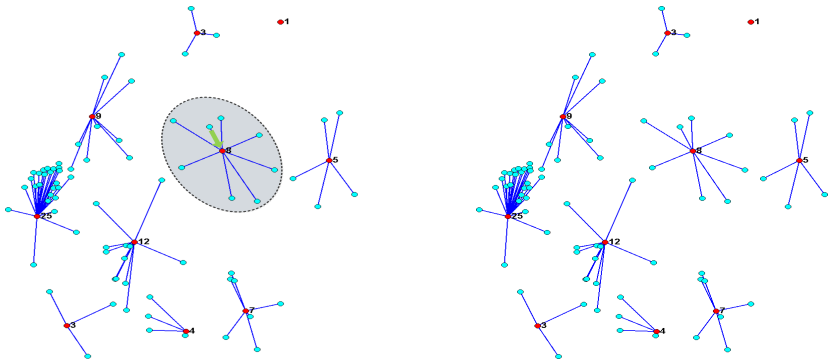
Abbildung 5.40: Iterative Optimierung des Algorithmus Dragoon mit der Nebenbedingung freie Platzierung



(a) Platzierung am Ende der Initialisierung

(b) Optimierung des erstens Zentrums in der ersten Iteration

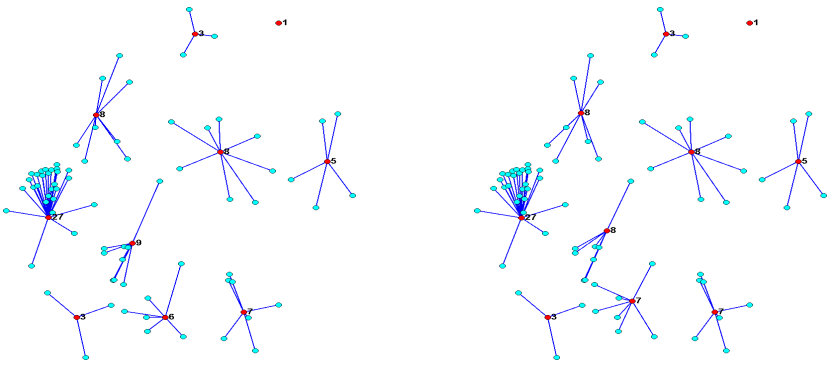
(c) Optimierung des zweiten Zentrums in der ersten Iteration



(d) Optimierung des dritten Zentrums in der ersten Iteration

(e) Optimierung aller Zentren nach der ersten Iteration

Abbildung 5.41: Erste Iteration der Optimierung von Dragoon mit der Nebenbedingung freie Platzierung



(a) Optimierung der Zentren nach der zweiten Iteration

(b) Ergebnis am Ende der Optimierung

Abbildung 5.42: Optimierung mit Dragoon mit der Nebenbedingung freie Platzierung

5.3.5 Spezialfall Backtrack

Auf Grundlage von Dragoon, der Idee einer erneuten Platzierung beim Greedy Verfahren und den Kenntnissen über die Problemstellung, entsteht ein weiterer Optimierungsalgorithmus Backtrack, der ausschließlich spezifisch für die Nebenbedingung der knotenbasierten Platzierung anwendbar ist. Dieser besteht wiederum aus einer Initialisierungsphase und einer Optimierungsphase. Zur Initialisierung verwendet der Algorithmus Backtrack den selben Initialisierungsprozess wie Dragoon. Der Unterschied besteht in der anschließenden Optimierung der initialen Lösung.

Zu Beginn der Optimierungsphase wird für alle Nutzer die Entfernung zu allen möglichen Positionen von Zentren berechnet und in einer Entfernungsmatrix abgespeichert. Anschließend startet ein iterativer Prozess, welcher die Zentren nacheinander aus der gegenwärtigen Lösung entfernt und durch ein neues Zentrum an einer optimierten Position ersetzt. Dazu werden die Zentren in der Reihenfolge betrachtet, in der diese dem Szenario hinzugefügt wurden.

Zu Beginn eines Optimierungsschrittes wird das jeweilige betrachtete Zentrum aus der Lösungsmenge entfernt. Anschließend erfolgt mittels der berechneten Matrix die Kalkulation des gegenwärtigen Zustandes für das Gesamtsystem. Dazu wird für jeden Nutzer die Entfernung zu seinem nächstgelegenen Zentrum aus den verbleibenden Zentren bestimmt. In Verbindung mit der Entfernungsmatrix vergleicht der Algorithmus anhand des Zielfunktionswertes für jeden Nutzer die Entfernung zum nächstgelegenen Zentrum mit der Entfernung zu allen potentiellen Standorten für ein weiteres Zentrum und speichert den besten Wert ab. Somit wird für jeden potentiellen Standort der Zielfunktionswert ermittelt, wenn dieser für ein neues Zentrum dient. Zum Schluss eines Optimierungsschrittes wird das entfernte Zentrum durch ein neues Zentrum an dem Standort ersetzt, welcher für das gesamte Szenario den besten Zielfunktionswert erreicht. Dieser Optimierungsprozess wird solange wiederholt, bis sich die Positionen der Zentren nicht mehr ändern oder die maximale Anzahl an Iterationen erreicht ist.

Die Begrenzung der Iteration ist notwendig, um die Terminierung sicher zu stellen, da keine Konvergenz garantiert ist. Für den Algorithmus ergibt sich somit ein quadratischer Berechnungsaufwand, wobei durch die Initialisierung eine zwei-approximierte Güte der Lösung erreicht wird.

6 Tragfähigkeitsnachweis und Evaluation

Zum Nachweis der Realisierbarkeit und der Funktionsweise erfolgt eine Implementierung des Konzeptes mittels der Programmiersprache Java. Zur nutzbringenden Anwendbarkeit des Modells für mehrere Bereiche werden verschiedene Optimierungen und Simulationen anhand von realistischen und synthetischen Szenarien durchgeführt. Dies dient zur Demonstration und Evaluierung des entstandenen Systems als auch der Analyse der Sachverhalte im Allgemeinen. Hierbei werden ausgewählte Optimierungsansätze miteinander verglichen. Weiterhin erfolgt hinsichtlich der Platzierung eine Gegenüberstellung der bestehenden Verfahren mit den neu entwickelten Algorithmen. Als Erstes sind die verschiedenen Testszenarien sowie der Optimierungs- und Simulationsumgebung beschrieben. Anschließend werden domänenübergreifende Aspekte analysiert bevor spezifische Sachverhalte dargestellt sind. Die ausgewählten Analysen stellen dabei Beispiele für den Einsatzzweck dar, wobei dieser nicht darauf begrenzt ist.

6.1 Implementierung

Dieser Abschnitt beschreibt die Umsetzung des in Kapitel 5 konzipierten Modells und der verschiedenen Lösungsansätze in Software. Dazu erfolgt eine Betrachtung der wesentlichen Komponenten und Zusammenhänge.

Die Abbildungen 6.1 bis 6.4 zeigen abstrakte UML-Klassendiagramme [307], welche einen Überblick des in Java implementierten Analysewerkzeugs geben. Auf Grund der Übersichtlichkeit sind dabei nur die wesentlichen Komponenten mit den grundlegenden Funktionen und Schnittstellen abgebildet.

In der Abbildung 6.1 ist der generische Ansatz zur Erstellung der Optimierungsaufgaben aus einem Szenario aufgezeigt. Ein Szenario besteht dabei aus den Nutzern mit deren Standorten und Profilen. Zudem sind gegebenenfalls explizit eine Infrastruktur, mehrere Quellen und bereits vorhandene Zentren vorgegeben. Der Generator erstellt aus dieser Problemstellung entsprechend der definierten Konfiguration Optimierungsaufgaben. Hierbei werden mit den Zentren der Optimierungsaufgabe unter anderem bereits Strukturen für die Lösung angelegt. Es gibt dabei entsprechend der Konfiguration die definierte Anzahl an Zentren mit zugehöriger Kapazität des Speichers und Speicherstrategie. Auf Grundlage der generischen Struktur werden die Daten entsprechend aufbereitet und zu einem Modell des zu analysierenden Sachverhaltes kombiniert. Das Szenario wird um die Informationen der zu verwendenden Algorithmen, Strategien und Funktionen angereichert. Dazu verwendet der Generator die vier verschiedenen Fabriken in Abbildung 6.2, um die notwendigen Verfahren zu erzeugen und die angegebenen Objekte zu erhalten.

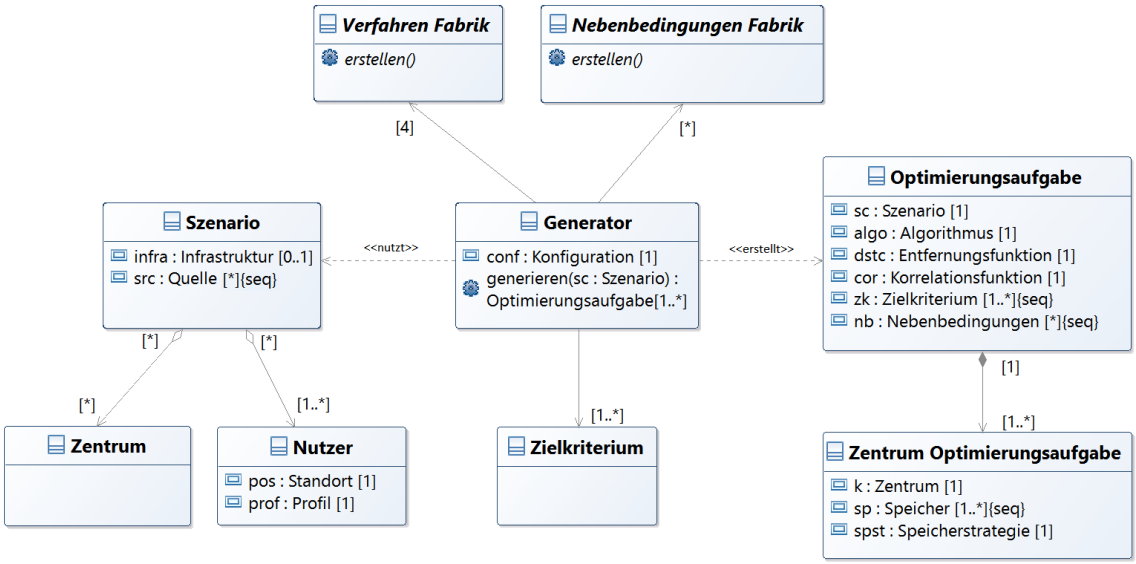


Abbildung 6.1: Abstraktes UML-Klassendiagramm von den wesentlichen Bestandteilen des generischen Modells

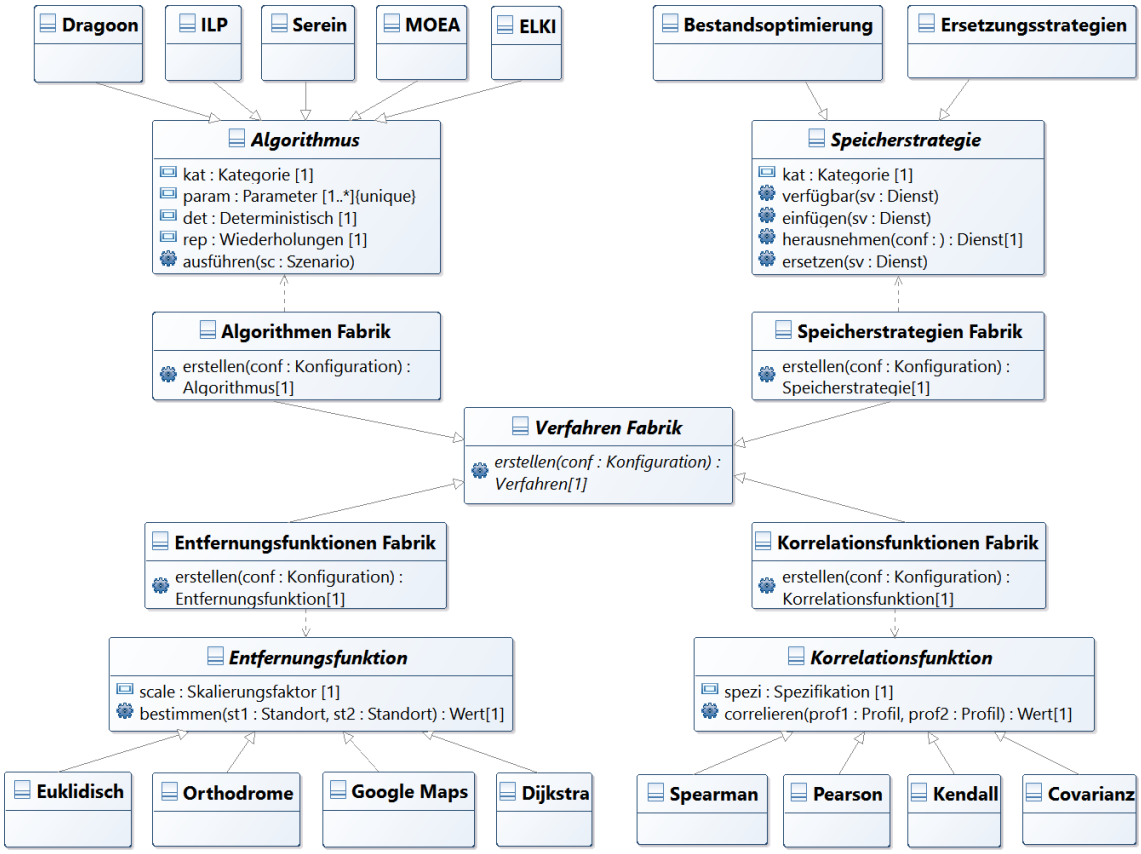


Abbildung 6.2: Abstraktes UML-Klassendiagramm von den verschiedenen Fabrikmethoden zur Erzeugung der Algorithmen, Strategien und Funktionen mittels Generator (Auf Grund besserer Übersichtlichkeit weicht die Darstellung des Design Patterns der Fabrikmethode vom klassischen Entwurfsmuster gemäß *Gang of Four* [308] ab.)

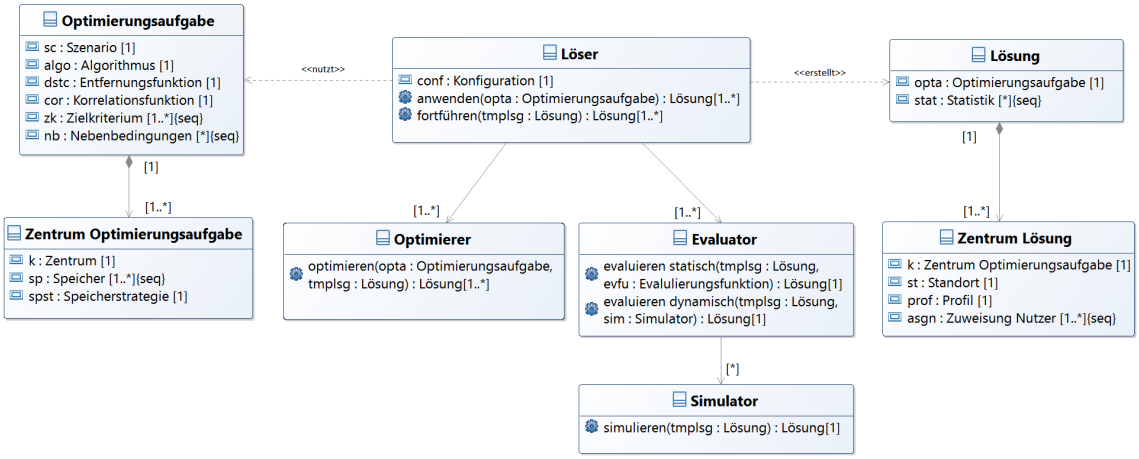


Abbildung 6.3: Abstraktes UML-Klassendiagramm von den wesentlichen Bestandteilen der Lösungsberechnung zu einem Optimierungsproblem

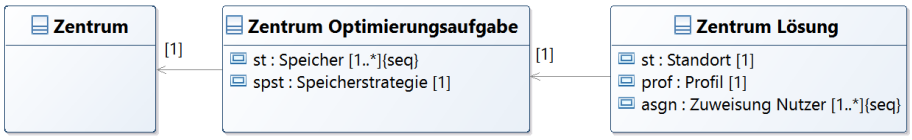


Abbildung 6.4: Abstraktes UML-Klassendiagramm vom Zentrum mit sukzessive erweiterten Informationen über die verschiedenen Bearbeitungsschritte hinweg

Hierbei sind die Algorithmen, Strategien und Funktionen in entsprechende Kategorien deren Anwendbarkeit eingeordnet. Darüber erfolgt die Anbindung weiterer Frameworks, welche bereits Implementierungen verschiedener Verfahren liefern. Zudem werden über den Generator entsprechend der Konfiguration die Optimierungsziele spezifiziert. Über eine separate Fabrik lassen sich Nebenbedingungen mit in die Optimierungsaufgabe einbeziehen.

In der Abbildung 6.3 sind die wesentlichen Klassen der simulationsbasierten Optimierung aufgezeigt. Aus der generierten Optimierungsaufgabe berechnet die Klasse Löser auf Grundlage des generischen Modells Ergebnisse und gibt diese entsprechend als Lösungen aus. Über die Konfiguration erfolgt dabei die Steuerung der Prozesse. Zunächst werden die Optimierer auf die Aufgaben angewendet und initiale sowie temporäre Zwischenlösungen erstellt. Diese lassen sich entsprechend fortführend weiter verarbeiten und optimieren. Zur Evaluation der Lösungen stehen statische und dynamische Methoden zur Verfügung. Insbesondere bei der dynamischen Evaluation kommt dabei ein Simulator zum Einsatz, welcher entsprechend der vorgegeben Konfiguration des Lösers vorgeht. Aus den Zwischenlösungen wird anschließend die Lösung erstellt und zurück gegeben. Diese beinhaltet neben Statistiken alle notwendigen Informationen für eine vollständige Lösung. Insbesondere gehören dazu die Zentren mit dem Standort und dem Profil sowie der Zuweisung zu bedienenden Nutzer.

Der Zusammenhang des Systems wird anhand der zentralen Komponente Zentrum ersichtlich. Dazu zeigt die Abbildung 6.4 den Übergang von der Problemstellung durch das Szenario über die Aufbereitung als Optimierungsaufgabe bis hin zur Erstellung der Lösung. Dabei werden der Klasse Zentrum sukzessiv Eigenschaften und Informationen hinzugefügt.

Zur Reduktion des Berechnungsaufwandes wird in allen Experimenten bei der Infrastruktur von einem ungerichteten Graphen ausgegangen, sofern nichts anderes definiert ist. Demzufolge gibt es ausschließlich bidirektionale Verbindungen. Damit verringert sich die Anzahl von Anfragen an externe Kartendienste und die Berechnungen zur Wegfindung mittels Dijkstra-Algorithmus.

6.2 Testdatensatz der Standorte und Infrastrukturen

Die Testdaten für Szenarien zur Standortplanung unterteilen sich in die drei Kategorien: Synthetisch, FLP und CDN. Dazu gibt die jeweilige Tabelle 6.1, 6.2 und 6.3 eine Übersicht der Daten. Zu jedem Szenario ist in Anhang D eine Abbildung gegeben, welche den Sachverhalt visualisiert. Nichtsdestotrotz ist der Ansatz auf beliebige Szenarien und Graphenstrukturen anwendbar, sofern diese dem Anspruch der Vollständigkeit genügen und keine negativen Kantengewichte besitzen. Dies ist bei realistischen Szenarien stets gegeben. Die Dreiecksungleichung 2.1 muss für die Berechnung einer Lösung nicht erfüllt sein, es ist lediglich bei der Lösungsfindung und Entfernungsermittlung darauf zu achten.

Um eine breite Basis von unabhängigen Szenarien zu erhalten, wurden zur Erzeugung der Daten verschiedene Generatoren eingesetzt. Bei den synthetischen Szenarien erfolgte die Standortgenerierung von den Nutzern zum Teil mittels dem Java internen Zufallsgenerator, wobei die Wertebereiche entsprechend für Geokoordinaten begrenzt sind. Weiterhin entstanden vier der synthetischen Szenarien über das Generierungswerkzeug Gephi [309]. Dieses nutzt dazu das Wissen aus den wissenschaftlichen Projekten von Generatoren für Graphen und Netztopologien von unter anderem Gt-Itm [310], Inet [311], BRITE [312] und TIERS [313]. Diese haben den Anspruch realistische Szenarien für Infrastrukturen zu erzeugen, werden jedoch nicht mehr weiterentwickelt. Darüber hinaus bietet die Community um das Framework ELKI noch zwei synthetische Szenarien an [314], welche ebenfalls zu Testzwecken verwendet werden.

Tabelle 6.1: Testszenarien der Kategorie Synthetisch

Index	Name	Knoten
1	Zufall Rechteck 1	100
2	Zufall Rechteck 2	100
3	Zufall Rechteck 3	200
4	Zufall Quadrat	200
5	Gephi 1	100
6	Gephi 2	200
7	Gephi 3	200
8	Gephi 4	300
9	ELKI density	150
10	ELKI Mouse	500

Die Szenarien der Logistik entsprechen den größten Städten in den angegebenen Ländern gemessen an der Einwohnerzahl [315–324]. Der exakte Standort mit den Geokoordinaten ergibt sich aus der Abfrage über Google Maps [238].

Tabelle 6.2: Testszenarien der Kategorie Logistik gemäß FLP

Index	Name	Knoten
11	Tschechien	62
12	Niederlande	67
13	Deutschland	77
14	Europa	98
15	Frankreich	114
16	Schweiz	144
17	Italien	147
18	Vereinigtes Königreich	196
19	Indien	230
20	United States of America	273

Der Datensatz für reale Netztopologien nutzt die Informationen des wissenschaftlichen Projektes *The Internet Topology Zoo* [19]. Hierbei sind neben den gegebenen Standorten die Verbindungsnetze bei der Wegfindung zu beachten. Bei diesen Szenarien werden neben der Eigenschaft der Knotenanzahl noch weitere Parameter mit angegeben, um die Komplexität zu verdeutlichen. Die verschiedenen Parameter Exzentrizität, Radius und Diameter sowie Entferntheit sind nachfolgend erläutert.

Tabelle 6.3: Testszenarien der Kategorie Rechnernetz gemäß CDN

Index	Name	Knoten	Kanten	$\varnothing$ *rad*	$\varnothing$ *dia*	$\varnothing$ *fer*	max *dia*
21	Latnet	68	73	4.4	4.8	4.0	12
22	Uninett	74	102	2.9	6.1	4.6	9
23	Dialtelecom (Cz)	124	126	10.4	23.8	14.6	39
24	GtsCe	141	181	6.4	13.9	9.4	24
25	Tata (Nld)	143	181	7.3	15.4	9.9	28
26	Colt	146	164	5.6	11.7	8.5	20
27	US Carrier	151	174	9.1	17.6	13	33
28	Cogentco	180	210	7.7	19	11.6	30
29	Westeuropa	271	385	4.2	11.5	7.0	18
30	Europa	447	605	4.6	12.0	7.7	21

Exzentrizität - $\varepsilon(v)$:
Entspricht der größten Entfernung zwischen einem Knoten v und dem am weitesten entfernt gelegenen Knoten gemäß der kürzesten Verbindungsstrecke, siehe Gleichung 6.1.

$$\varepsilon(v) = \max \forall v \in V\{\min d(v_i, v_j)\} \tag{6.1}$$

Radius - *rad*:
Entspricht dem minimalen Wert der Exzentrizität für einen beliebigen Knoten des Graphen, siehe Gleichung 6.2.

$$rad = \min_{v \in V} \varepsilon(v) \tag{6.2}$$

Diameter - *dia*:
Entspricht dem maximalen Wert der Exzentrizität für einen beliebigen Knoten des Graphen, siehe Gleichung 6.3.

$$dia = \max_{v \in V} \varepsilon(v) \tag{6.3}$$

Entferntheit / Verbundene Zentralität - *fer*:
Beschreibt die durchschnittliche Entfernung zwischen allen Knotenpaaren bei der Verwendung des kürzesten Pfades, siehe Gleichung 6.4.

$$fer = \frac{\sum_{v_i=1, i \neq j}^{n} d(v_i; v_j)}{n-1} \tag{6.4}$$

Bei allen Szenarien wird während der Optimierung von einer symmetrischen und bidirektionalen Infrastruktur ausgegangen, um die Berechnungsaufwand der Entfernungsfunktion zu verringern. Weiterhin besitzen alle Verbindungen das gleiche Gewicht, sofern nicht anders angegeben, sodass sich die Entfernung auf Grund Längenmetrik ergibt, um die Ergebnisse insbesondere bei der grafischen Darstellung nachvollziehbar zu halten. Auf Grund der gleichmäßigeren Verteilung der Knoten bei den synthetischen Testszenarien als bei den realen Datensätzen gestaltet sich die Lösungsfindung bei diesen Problemstellungen komplizierter. Daher ist es für die Optimierungsalgorithmen schwieriger, eine optimierte Lösung zu finden, wodurch sich diese Kategorie für Leistungstests besonders eignet. Zudem sind bei der Nutzung der euklidischen Entfernungsfunktion und direkten Verbindungen keine aufwendigen Berechnungen zur Wegfindung durchzuführen.

6.3 Testdatensatz der Profile von Nutzern

Die Profile sind für die Tests entsprechend der Beschreibung in Kapitel 5.1.4 umgesetzt, wobei diese sich zur besseren manuellen Nachvollziehbarkeit der Ergebnisse nicht zeitabhängig verändern. Jeder Nutzer erhält ein individuelles Profil bei der Initialisierung des Szenarios.

Die Testdaten für die Anfrageprofile von Nutzern unterteilen sich in die zwei Kategorien synthetische Zipf-Verteilung und reale Traces. Bei der synthetischen Zipf-Verteilung wird für den Parameter α der realistische Wertebereich von 0,6 bis 2 gemäß den in Kapitel 4.5 durchgeführten Analysen angewendet. Sofern nicht abweichend deklariert, besteht das Universum aus 100 verschiedenen Diensten, welche in den Profilen auftreten und an den Zentren sich anfragen lassen. Beim Einsatz von realen Traces werden diese entweder

direkt durch die entsprechende Ausführung der ausgezeichneten Reihenfolge benutzt oder über die Umsetzung als Profil mit zufälliger Auswahl gebraucht. Traces lassen sich dabei in fünf verschiedene Zugriffspattern einteilen, wobei eine Zuordnung bekannter Traces erfolgt [325]:

- Sequentielle Referenzen
- Schleifen Referenzen: CS, Glimpse, Postgres
- Zeitlich gruppierte Referenzen: Spirite
- Wahrscheinlichkeitsbasierte Referenzen: Zipf, CPP, 2-Pools
- Mixed: Multi 1-3

Im Folgenden wird eine Beschreibung der Traces gegeben, welche hier Anwendung finden:

- 2-Pools: Synthetisch erstellter Trace mit 100.000 Einträgen, der Anwendungsverhalten nachbildet [271]
- CPP: Trace des GNU C Preprozessor Compilers unter Eingabe des Quellcodes eines ca. 11 MB großen Programms [325, 326]
- CS: Trace eines interaktiven C Quellprogramms eines Überprüfungswerkzeuges bei Eingabe eines ca. 9 MB großen Quellcodes [325, 326]
- Glimpse: Basierend auf einem Text-Informationsgewinnung-Werkzeug erstellter Trace (ca. 50 MB) [325, 326]
- Sprite: Referenzierte Blocknummern einer Festplatte des netzbasierten Dateisystems eines zentralen Dateiservers bei Nutzeranfragen über einen Zeitraum von zwei Tagen [327]
- Multi 1: Aufgezeichneter Trace bei der parallelen Ausführung von CPP und CS [225]
- Multi 2: Aufgezeichneter Trace bei der parallelen Ausführung von CPP, CS und Postgres [225] (Der Postgres Trace basiert auf Anfragen an eine relationale Datenbank von der Universität von Californien in Berkley.)
- Multi 3: Aufgezeichneter Trace bei der parallelen Ausführung von CPP, Gnuplot, Glimpse und Postgres [225]

In der Gesamtheit dienen die Profile maßgeblich zum Vergleich der verschiedenen Speicherverwaltungsstrategien. Darüber hinaus lassen sich Rückschlüsse auf die Platzierung der Zentren als auch die Zuweisung der Nutzeranfragen ziehen.

6.4 Optimierungs- und Simulationsumgebung

Zur Analyse und Auswertung der Ergebnisse werden die Daten der einzelnen Szenarien entsprechend akkumuliert und der Durchschnitt gebildet, sofern ein Experiment nicht spezifische Effekte für einzelne Szenarien untersucht. Die Entfernungen zwischen den Knoten werden normiert, um bei kurzen und langen Strecken von kleinen und großen Szenarien den Einfluss auszugleichen und eine Vereinheitlichung über die verschiedenen Szenarien zu erreichen. Dazu erfolgt die Angabe der Entfernung zwischen zwei Knoten

in prozentualer Relation zur maximalen Entfernung zwischen den zwei am weitesten entfernten Knoten gemäß der kürzesten Verbindungsstrecke des jeweiligen Szenarios. Die Simulation der Dienste wird mit der einheitlichen Größe von eins durchgeführt, da die gängigen Speicherverwaltungsstrategien darauf ausgelegt sind und um einen Vergleich zu den bestehenden Arbeiten aus Abschnitt 4.5 ziehen zu können. Diese arbeiten ebenfalls vorrangig mit einer Einheitsgröße. Für Dienste unterschiedlicher Größe lassen sich die Speicherverwaltungsstrategien ebenfalls anwenden, wobei zur Leistungssteigerung gegebenenfalls Anpassungen vorzunehmen sind. Als Testdatensatz wird die Vereinigungsmenge der drei Kategorien Synthetisch, FLP und CDN zur Optimierung und Simulation verwendet, sofern kein spezifischer oder domänenabhängiger Aspekt untersucht wird.

6.5 Evaluierung der Platzierungsleistung

Zu Beginn der Evaluierung erfolgt eine Analyse zur Platzierungsleistung des neu entwickelten Algorithmus Dragoon in Bezug zu den bestehenden Ansätzen. Dies geschieht anhand verschiedener Optimierungskriterien und Bewertungsfunktionen des generischen Modells. Hierbei werden bereits verschiedene Eigenschaften des generischen Modells aufgezeigt und nachgewiesen.

6.5.1 Allgemeiner Vergleich der Algorithmen bei unterschiedlichen Anwendungsgebieten

Um den Einfluss unterschiedlicher Szenarien und Entfernungsfunktionen auf die Optimierungsleistung der Algorithmen auswerten zu können, erfolgen in diesem Abschnitt zwei Analysen. Analyse 1 basiert auf dem Mittelwert mehrerer Szenarien, wohingegen Analyse 2 ein spezifisches Szenario betrachtet. Dabei werden jeweils die Szenarien der drei Kategorien Logistik, Rechnernetze und Synthetisch betrachtet. Als Nebenbedingung ist die Platzierung am Knoten gefordert, da dies bei allen Szenarien und Algorithmen möglich ist und einen übergreifenden Vergleich zulässt. Die Optimierung erfolgt hinsichtlich der Minimierung der maximalen Entfernung eines Nutzers zu seinem nächstgelegenen Zentrum.

Analyse 1:
In den folgenden drei Abbildungen 6.5, 6.6 und 6.7 sind Leistungsvergleiche von mehreren Algorithmen dargestellt. Die Platzierung der Zentren erfolgt dabei anhand von den Szenarien des jeweiligen Anwendungsbereiches mit den zugehörigen Entfernungsfunktionen.

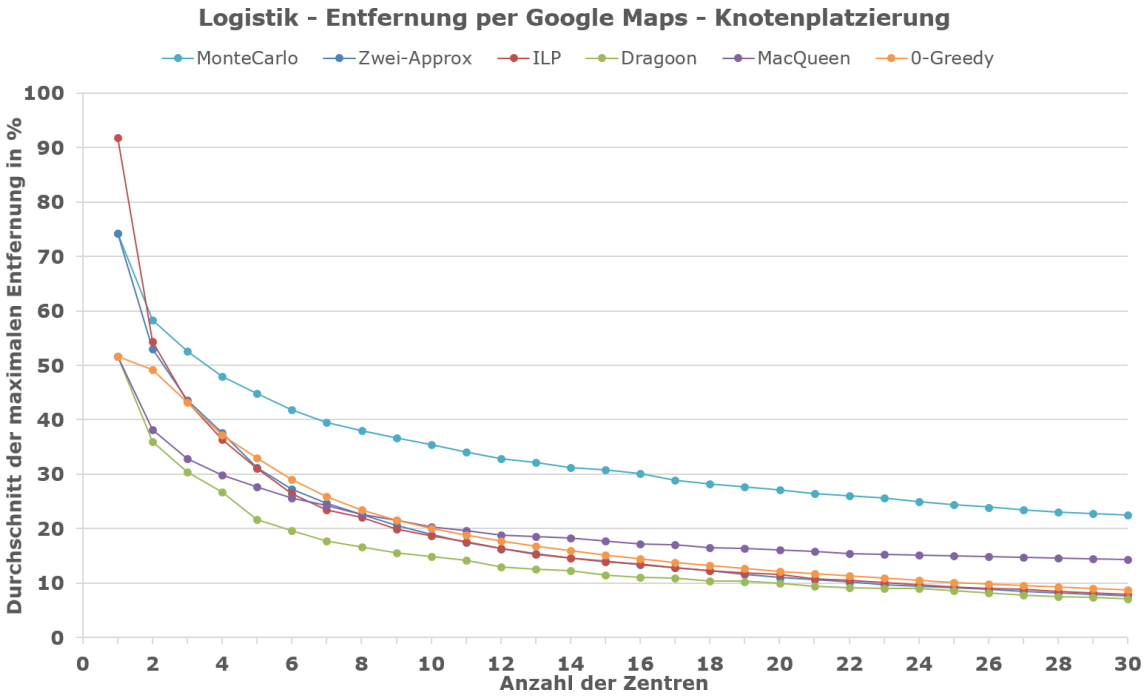


Abbildung 6.5: Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand der Testszenarien Logistik mit der Entfernungsberechnung gemäß Google Maps (Kleinere Werte sind besser.)

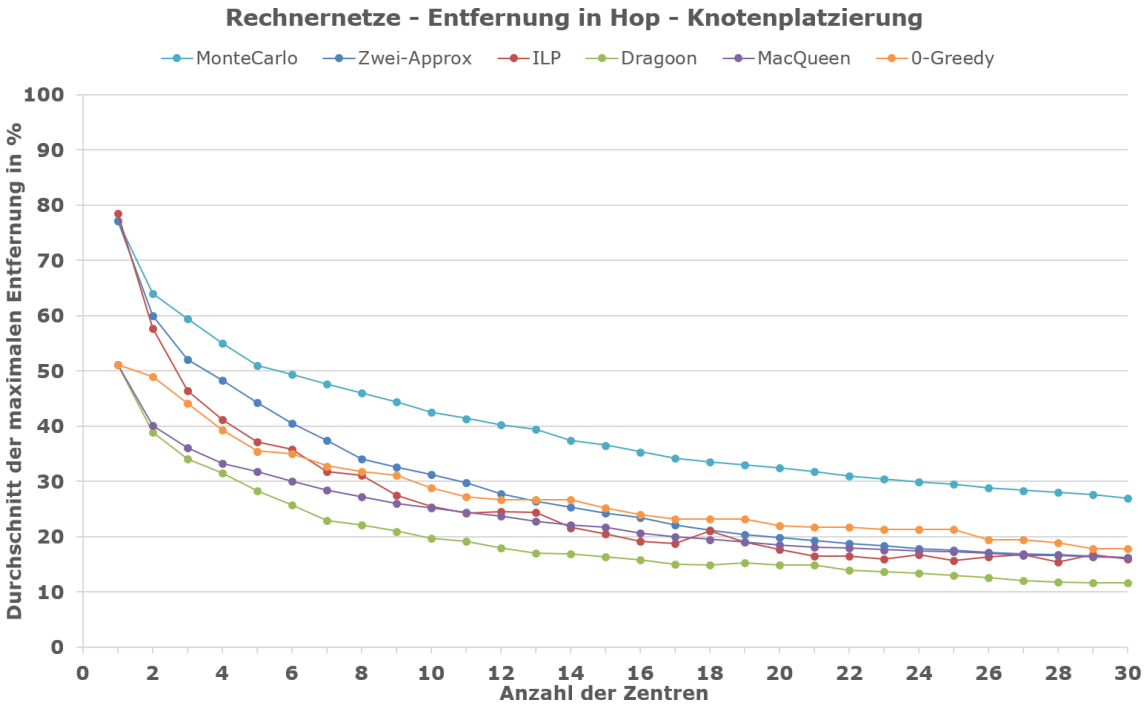


Abbildung 6.6: Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand der Testszenarien Rechnernetze mit der Entfernungsberechnung über Dijkstra auf dem Graphen (Kleinere Werte sind besser.)

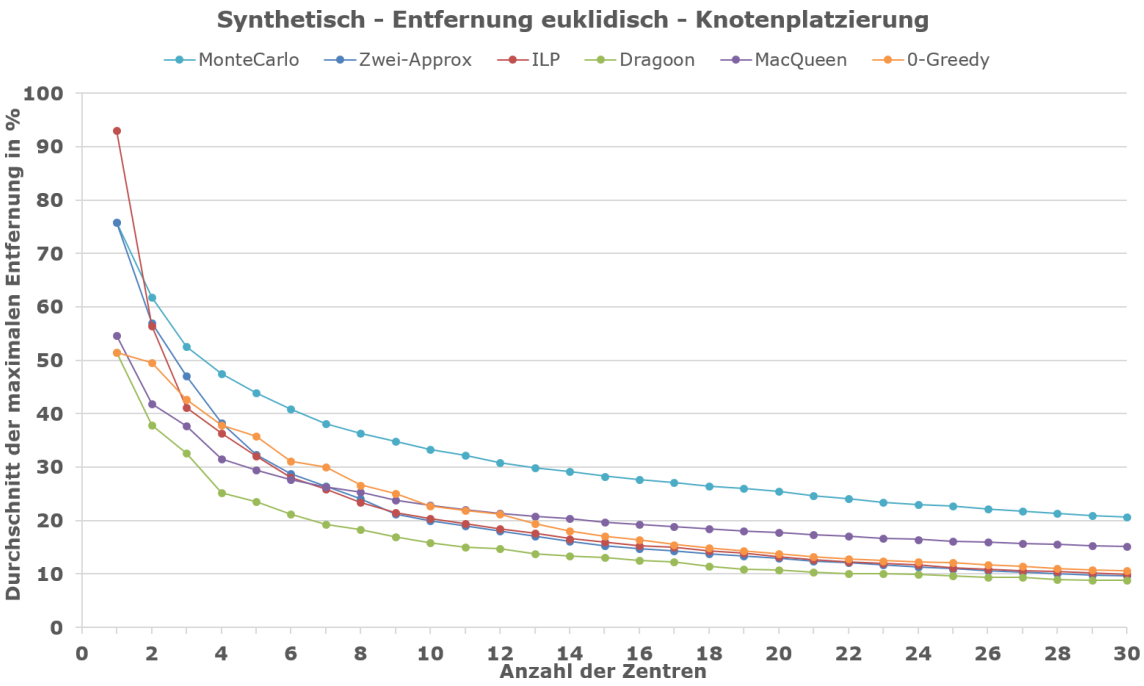


Abbildung 6.7: Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand der Testszenarien Synthetisch mit der Entfernungsberechnung gemäß Euklid (Kleinere Werte sind besser.)

Auswertung:
Die Platzierungsleistung eines Algorithmus ist von dem jeweiligen Anwendungsgebiet und dem einzelnen Szenario abhängig, sodass die Ergebnisse entsprechend variieren. Im Vergleich eines Algorithmus zu den anderen Optimierungsverfahren ist das Verhältnis über die verschiedenen Anwendungsbereiche annähernd affin. Trotz der unterschiedlichen Szenarien und den verschiedenen Funktionen zur Berechnung der Entfernung ergibt sich bei den Algorithmen im Mittel eine vergleichbare Rangfolge bezüglich der Platzierungsleistung über die Anzahl der Zentren hinweg.

Zum initialen Vergleich der Algorithmen dient als Referenz die Monte Carlo Methode. Dabei zeigt sich, dass die Ergebnisse des Zwei-Approx Algorithmus gegenüber Monte Carlo im Durchschnitt wesentlich besser und stabiler sind. Daher wird in den folgenden Abschnitten der Zwei-Approx Algorithmus auf Grund der zwei-approximierenden Güte als Referenz in Bezug gesetzt und die Monte Carlo Methode nicht weiter beachtet.

Analyse 2:
Zur weiteren Analyse wird in den folgenden Abbildungen 6.8, 6.9, 6.10 und 6.11 der Sachverhalt anhand des Testszenarios Dialtelecom explizit dargestellt. Dabei kommt die weitere Entfernungsfunktion gemäß Orthodrome zum Einsatz.

Auswertung:
Das zuvor beschriebene Verhalten wiederholt sich bei diesen Tests. Hierbei wird ersichtlich, dass Dragoon ein stabileres Verhalten im Bezug zu der Anzahl von Zentren hat als die Referenz-Algorithmen.

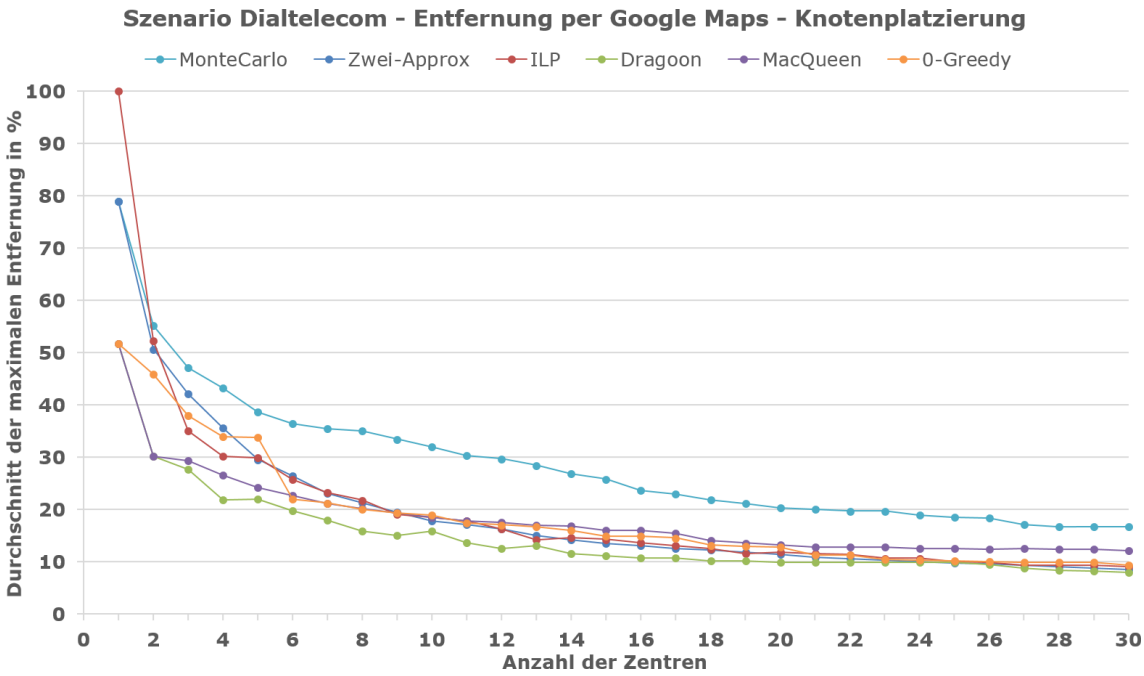


Abbildung 6.8: Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand des Szenarios Dialtelecom mit der Entfernungsberechnung gemäß Google Maps (Kleinere Werte sind besser.)

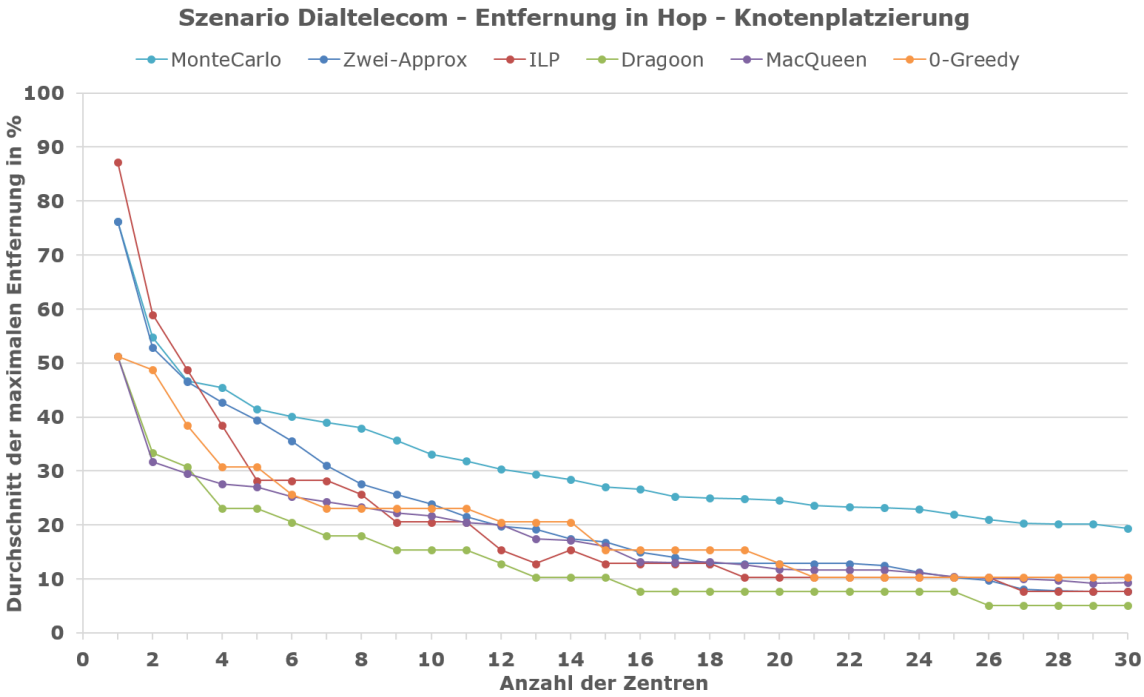


Abbildung 6.9: Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand des Szenarios Dialtelecom mit der Entfernungsberechnung über Dijkstra auf dem Graphen (Kleinere Werte sind besser.)

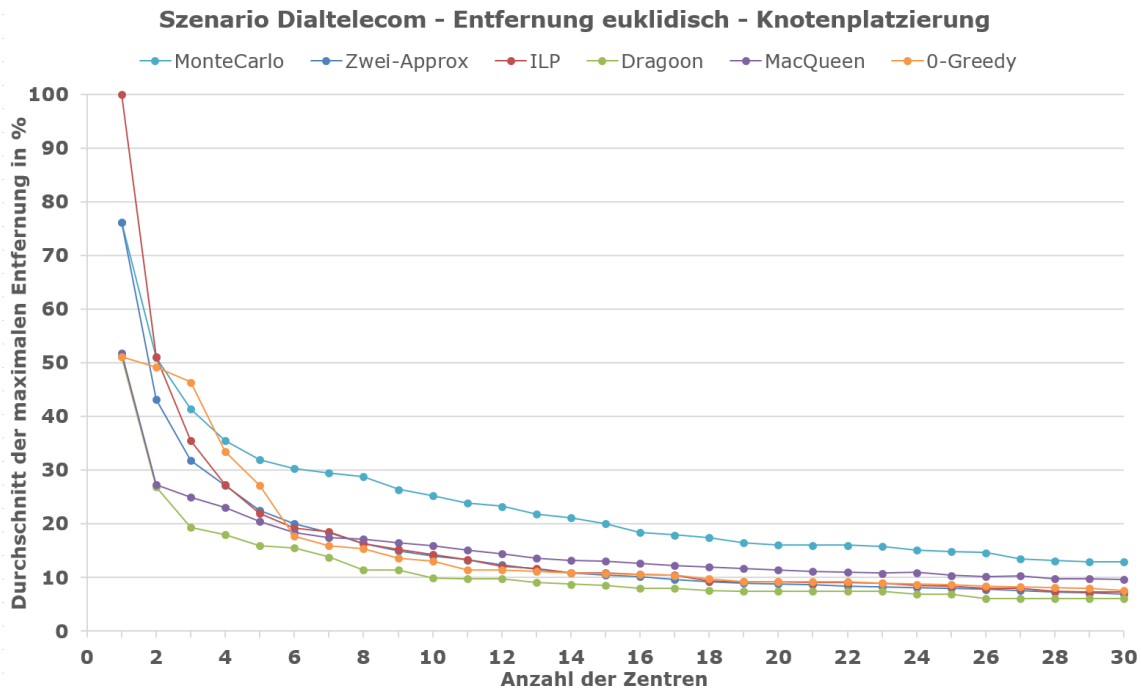


Abbildung 6.10: Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand des Szenarios Dialtelecom mit der Entfernungsberechnung gemäß Euklid (Kleinere Werte sind besser.)

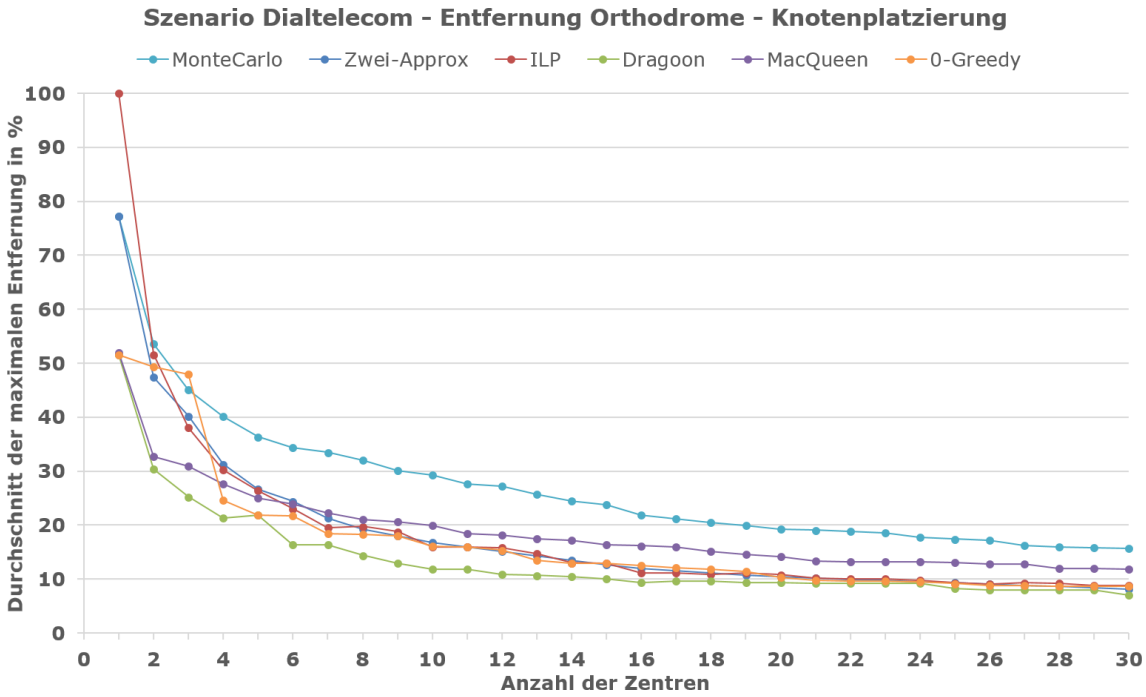


Abbildung 6.11: Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand des Szenarios Dialtelecom mit der Entfernungsberechnung gemäß Orthodrome (Kleinere Werte sind besser.)

6.5.2 Leistungsvergleich von Dragoon mit dem theoretischen Optimum

Basierend auf der garantierten Leistung des Zwei-Approx Algorithmus lässt sich gemäß der Gleichung 4.2 ein theoretischer Zielfunktionswert berechnen, der nahezu dem Wert bei optimaler Platzierung der Zentren entspricht. Auch wenn dabei keine Kenntnis über die konkrete Ausprägung der optimalen Lösung besteht, wird damit eine untere Schranke für den Zielfunktionswert des globalen Optimums definiert. Der jeweilige Zielfunktionswert des tatsächlichen Optimums bei optimaler Anordnung der Zentren hat den gleichen bzw. einen höheren Wert. Um das theoretische Optimum möglichst nahe dem realen Optimum anzugleichen, erfolgt eine wiederholte Ausführung des Zwei-Approx mit unterschiedlichen Zufallsentscheidungen. Aus den ermittelten Lösungen wird diejenige mit dem höchsten Zielfunktionswert zur Berechnung des theoretischen Optimums herangezogen, da das Gütekriterium des Algorithmus immer gilt. Die maximalen Zielfunktionswerte werden dazu entsprechend halbiert. Die ermittelten Leistungsdaten für das theoretische Optimum basieren auf dem Zwei-Approx Algorithmus mit jeweils 400 Ausführungen pro Szenario und Anzahl von Zentren. Eine Berechnung des tatsächlichen Optimums über einen Brute Force Ansatz ist im Vergleich wesentlich aufwändiger, sodass eine heuristische Berechnung über den Zwei-Approx Algorithmus erfolgt.

Die Abbildung 6.12 zeigt den Leistungsvergleich der verschiedenen Algorithmen. Die zum Vergleich genutzten Szenarien ist die Vereinigungsmenge aller Testdatensätze mit der euklidischen Entfernungsberechnung.

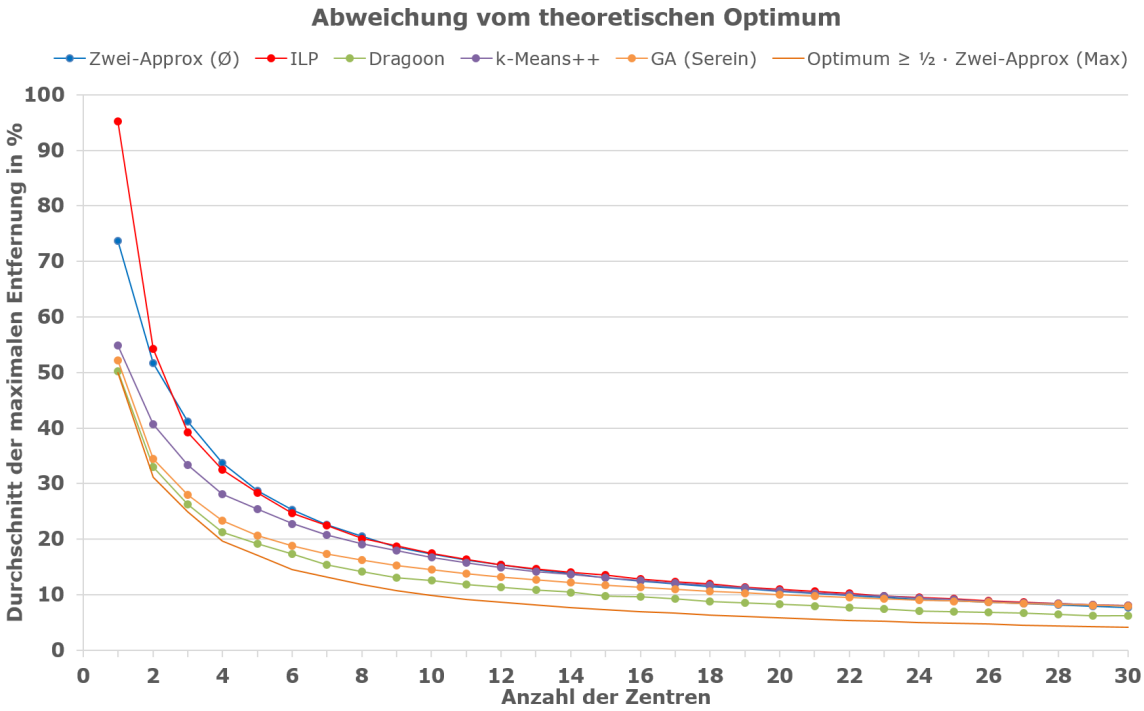


Abbildung 6.12: Abweichung der Leistung vom theoretischen Optimum hinsichtlich des Kriteriums der maximalen Entfernung (Kleinere Werte sind besser.)

Auswertung:
Der Vergleich verdeutlicht die Leistungsfähigkeit des entwickelten Algorithmus Dragoon, wobei die Ergebnisse der freien Platzierung genutzt werden. Bei einer geringen Anzahl an Zentren liegen die Ergebnisse von Dragoon nahe dem theoretischen Optimum. Mit steigender Anzahl von Zentren stagniert die Leistung aller Algorithmen auf einem annähernd gleichen Niveau. Dies zeigt, dass die Ergebnisse nahe dem tatsächlichen Optimum liegen. Zudem erreicht der Ansatz von Dragoon bessere Ergebnisse als Zwei-Approx und besitzt einen geringeren Rechenaufwand als ein Brute-Force Ansatz.

6.5.3 Einfluss von δ-Greedy auf die Platzierungsleistung im Vergleich zu Dragoon und Backtrack

Der Algorithmus Greedy lässt sich durch den Einsatz von Backtracking gemäß Kapitel 4.8.3 erweitern. Entsprechend der definierten Anzahl von δ Schritten, werden zuvor getroffene Entscheidungen überdacht. Der Algorithmus wird demzufolge als δ-Greedy bezeichnet. Mit jedem Schritt erhöht sich der benötigte Rechenaufwand nahezu linear um den Faktor δ+1 gegenüber 0-Greedy. In dieser Untersuchung wird die Auswirkung von Backtracking auf die Platzierungsleistung ermittelt und dem Ansatz von Dragoon gegenübergestellt. Insbesondere erfolgt ein Vergleich mit dem entwickelten Ansatz Backtrack, der von vornherein auf die Strategie von Backtracking ausgelegt ist. Die Abbildung 6.13 zeigt die Leistungsunterschiede ungeachtet des Berechnungsaufwandes.

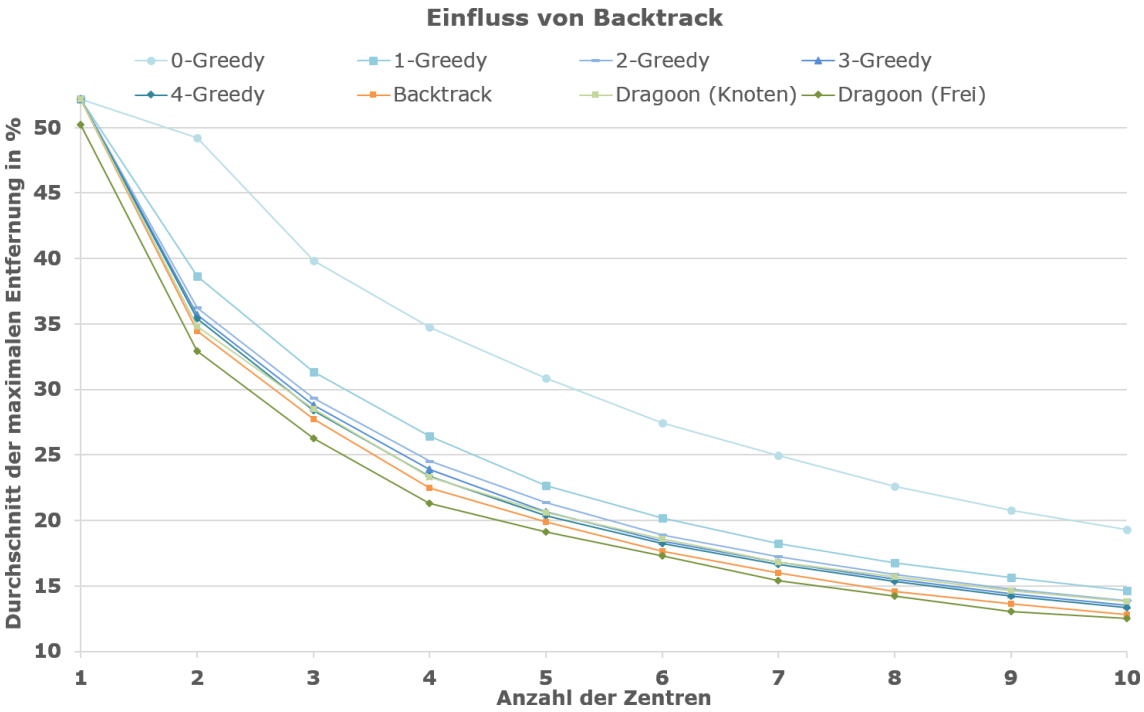


Abbildung 6.13: Einfluss von Backtracking auf die Platzierungsleistung ungeachtet des steigenden Berechnungsaufwandes (Kleinere Werte sind besser.)

Auswertung:
Mit steigender Anzahl von δ-Backtracking Schritten beim Algorithmus Greedy stagniert die Leistungsverbesserung. Zudem verringert sich mit steigender Anzahl von Zentren die Leistungsunterschied zwischen den Backtracking-Schritten. Daher ist der erhöhe Berechnungsaufwand nicht mehr gerechtfertigt, wodurch ein alternativer Ansatz zu bevorzugen ist. Darüber hinaus wird ersichtlich, dass die beiden Optimierungsansätze Dragoon und Backtrack im Durchschnitt bessere Ergebnisse erreichen. Insbesondere gelingt es Dragoon ohne den Einsatz von Backtracking bessere Ergebnisse als 1-Greedy zu erzielen. Nichtsdestotrotz eröffnet sich dadurch weiteres Potential, um die Leistung des Algorithmus Dragoon von Kapitel 5.3 weiter zu steigern. Dafür gibt es eine erweitere Variante des Algorithmus, wobei nach dem Optimierungszyklus ein Backtracking-Schritt ausgeführt wird. Dieser erfolgt ebenfalls nach dem bisherigen Prinzip der Optimierung. Der Unterschied besteht darin, dass sich in jeder Iteration abwechselnd die Reihenfolge, in der die Zentren betrachtet werden, entsprechend der längsten und kürzesten Strecke zu einem Nutzer verändert.

6.5.4 Überprüfung der Robustheit anhand von Leistungsdifferenzen bei der Platzierung

Robuste und schnelle Lösungsansätze sind insbesondere bei zeitkritischen Einsatzzwecken erforderlich. Zur Bestätigung der Platzierungsleistung des Algorithmus Dragoon erfolgt eine Evaluierung anhand der Extrembedingung möglichst ungünstiger Szenarien für den neu entwickelten Ansatz im Vergleich zu den anderen Optimierungsverfahren. Dazu werden über das evolutionäre Optimierungsverfahren des genetischen Algorithmus Instanzen des Problems erzeugt, welche besonders herausfordernd für Dragoon sind und bei denen ein alternativer Algorithmus möglichst gute Lösungen erreicht. Ausgehend von einem Datensatz zufälliger Szenarien werden dazu über mehrere Iterationen hinweg diese durch den genetischen Algorithmus weiterentwickelt und verändert. Dazu erfolgt eine Verschiebung der Nutzerpositionen. Als Grundlage für den genetischen Algorithmus wird das Framework SEREIN aus Kapitel 4.8.7 verwendet. Zur Analyse nutzt der Algorithmus zur Reproduktion von Probleminstanzen eine Einzelpunkt Gauss-Mutation mit der Standardabweichung von 0,05 und einer arithmetischen Rekombination mit der Wahrscheinlichkeit von 0,3 [328]. Für die Auswahl kommt die Tournament Selektion mit der Größe von 2 zur Anwendung. Die Populationsgröße beträgt dabei 30 Instanzen über 120 Generationen.

In den Experimenten wird die relative Leistung von den zwei Optimierungsalgorithmen *Prüfling* und *Herausforderer* für ein Szenario gemäß der Gleichung 6.5 bestimmt. Bei der Optimierung über den genetischen Algorithmus besteht dabei das Ziel eine möglichst große Leistungsdifferenz der zwei getesteten Algorithmen Herausforderer und Prüfling zu erreichen. Dazu wandelt der genetische Algorithmus die Szenarien so ab, dass der Herausforderer eine möglichst gute Leistung erreicht während dem Prüfling die Ermittlung einer vergleichbaren oder übertreffenden Lösung nicht möglich ist. Da das Framework SEREIN hinsichtlich einer Optimierung zum Minimum ausgelegt ist und das Ziel bei der Platzierung die Reduktion der Entfernung ist, bedeutet ein negativer Wert für ΔD, dass

der Herausforderungs-Algorithmus ein bessere Lösung erreicht hat als der Prüflings-Algorithmus.

$$\Delta D = D_{Herausforderungs-Algorithmus} - D_{Prüflings-Algorithmus} \tag{6.5}$$

Auf Grundlage der vorherigen Ergebnisse erfolgt die Analyse anhand der Knotenplatzierung mittels euklidischer Distanzberechnung. In den generierten Szenarien liegen die Koordinaten der Knoten im Bereich von 0 bis 100, wodurch eine quadratische Fläche entsteht. Die Tabelle 6.4 zeigt die Parameter für die Experimente. Es werden vorrangig Quadratzahlen für die Anzahl der Nutzer und der Zentren verwendet, da sich diese einfacher auf die quadratische Fläche der generierten Szenarien bei Gedankenexperimenten anwenden lässt.

Tabelle 6.4: Testdaten der Experimente

Experiment	I	II	III	IV	V	VI
Nutzer	10	25	36	49	49	64
Zentren	2	4	4	9	4	16

Zunächst erfolgt eine allgemeine Evaluation der Leistung von den Algorithmen bei 2000 zufällig generierten Szenarien. Die Tabelle 6.5 gibt dafür die durchschnittliche Leistungsabweichung ΔD_{avg} zwischen dem jeweiligen Algorithmus und Dragoon an. Anschließend erfolgt in der Tabelle 6.6 eine Gegenüberstellung der Ergebnisse mehrerer Algorithmenpaare mit der Optimierung von Szenarien hinsichtlich des Herausforderungs-Algorithmus.

Tabelle 6.5: Durchschnittliche Leistung bei der knotenbasierten Platzierung von alternativen Algorithmen im Vergleich zu Dragoon (ΔD_{avg}) basierend auf 2000 zufälligen Szenarien (Kleinere ΔD_{avg} Werte sind besser.)

Nutzer / Zentren	Zwei-Approx	MacQueen	Lloyd	0-Greedy	ILP	Backtrack
10 / 2	10,46	4,38	2,83	10,24	8,8	-0,10
25 / 4	7,5	5,21	3,85	6,52	4,99	-0,85
36 / 4	8,48	4,95	3,81	7,68	5,57	-0,74
49 / 4	8,97	4,61	3,6	8,09	5,85	-0,75
49 / 9	3,38	6,09	3,21	4,22	2,44	-1,05
64 / 4	9,59	4,34	3,41	8,54	6,08	-0,64
64 / 16	1,37	5,35	2,82	2,53	1,06	-1,17

Auswertung:
Zusammenfassend zeigt sich, dass es für jeden Optimierungsansatz ein Szenario gibt, bei dem ein anderer ein besseres Ergebnis liefert. Dennoch wird anhand der Werte ersichtlich, dass es mit den bisherigen Verfahren schwierig ist eine Probleminstanz zu finden, die gegenüber den neu entwickelten Algorithmen Dragoon und Backtrack eine bessere Lösung erreicht. In diesen Fällen ist es wichtig, solche Szenarien zu erkennen und das Potenzial eines Ausfalls über entsprechende Rückfallmaßnahmen zu beheben. Im Fall vom Algorithmus Dragoon wurden die Szenarien näher untersucht, bei denen

Tabelle 6.6: Gegenüberstellung der Δ D Leistung mehrerer Algorithmen am Beispiel von Problemstellungen mit 49 Nutzern und 4 Zentren (Kleinere Δ D Werte sind besser.)

Herausfor.	**Prüfling**						
	Zwei-Approx	MacQueen	Lloyd	0-Greedy	ILP	Dragoon	Backtrack
Zwei-Approx		-49,65	-25,18	-27,84	-18,73	-7,43	0
MacQueen	-24,9		-32,68	-25,36	-21,27	-9,49	-6,26
Lloyd	-24,37	-27,69		-25,98	-20,39	-9,81	-8,9
0-Greedy	-21,86	-23,61	-21,28		-18,5	-7,8	-9,62
ILP	-22,31	-34,24	-21,27	-29,47		-10,86	-1,57
Dragoon	-24,16	-50,05	-33,12	-25,61	-22,08		-7,58
Backtrack	-24,5	-25,7	-26,36	-33,54	-22,32	-10,85	

dieser eine vergleichsweise schwache Leistung erbringt. Dies führte zu der Erkenntnis, dass bei diesen einzelne Nutzer an entfernten, isolierten Positionen sind. Zudem zeigte eine schrittweise Betrachtung der Optimierung, dass sich insbesondere bei solchen Szenarien die Zentren bei Einhaltung der sequentiellen Auswahl in deren Verschiebung blockieren. In der Gesamtheit zeigt sich jedoch, dass der neu entwickelte Algorithmus im Durchschnitt eine bessere Leistung erbringt als die referenzierten Ansätze. Durch diese Analyse lassen sich Problemfälle ermitteln und die gewonnenen Erkenntnisse in das Design des Algorithmus einfließen. Dabei hat sich ergeben, dass ein Problemaspekt die Reihenfolge ist, in der die Zentren optimiert werden. In der erweiterten Variante Dragoon+ des Algorithmus Dragoon erfolgt eine Änderung der Reihenfolge nach jeder Iteration, indem diese stets invertiert wird. Dadurch lassen sich gegenseitige Blockierungen von Zentren auflösen.

6.5.5 Vergleich der Auswirkung von Platzierungsbedingungen

Dieser Abschnitt betrachtet die Unterschiede bei der Platzierung hinsichtlich der knotenbasierten Platzierung und der freien Platzierung. Dadurch lässt sich der Leistungsunterschied bei verschiedenen Nebenbedingungen für ein Gesamtsystem aufzeigen. Weiterhin werden hierbei die Unterschiede in der Komplexität ersichtlich als auch die Leistung der Algorithmen bei verschiedenen Bedingungen getestet. Die Analyse erfolgt unter Verwendung des Algorithmus Dragoon anhand aller Testszenarien mit der Entfernungsfunktion gemäß Euklid. Die Abbildung 6.14 zeigt die Entfernungsunterschiede bei der jeweiligen Anzahl an Zentren. Dabei erfolgte die Optimierung hinsichtlich des Kriteriums der maximalen (MAX) sowie der durchschnittlichen (AVG) Entfernung.

Die Tabelle 6.7 zeigt die statistischen Werte zu dem Unterschied bei den Platzierungsbedingungen Knotenplatzierung und freie Platzierung. Die Werte wurden auf Grundlage der Formel 6.6 berechnet.

$$\textit{Differenz}_{\text{Knoten-Frei}} = \frac{\text{Kriterium Knotenplatzierung}}{\text{Kriterium freie Platzierung}} \tag{6.6}$$

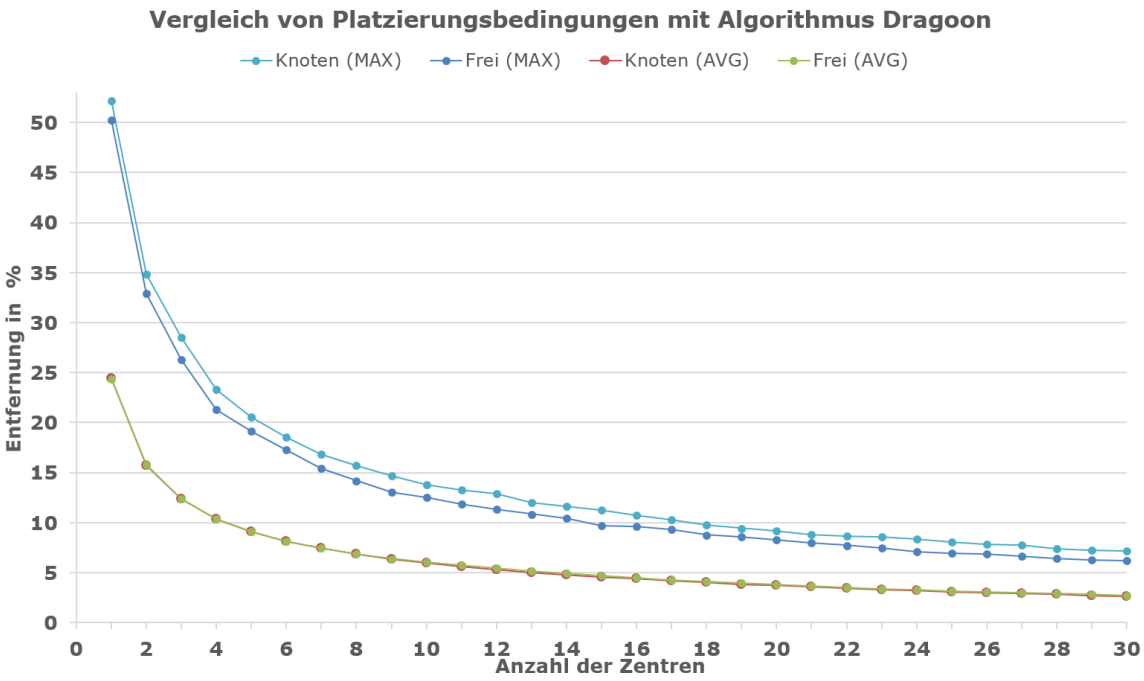


Abbildung 6.14: Vergleich der Optimierungsleistung bei der Platzierung von Zentren unter Beachtung der Nebenbedingung knotenbasierte und freie Platzierung (Kleinere Werte sind besser.)

Tabelle 6.7: Statistische Auswertung der Leistungsdifferenzen von der Platzierung bei den Nebenbedingungen Knotenplatzierung und freie Platzierung

Kriterium	Durchschnitt	Median	Maximum	Minimum
Maximale Entfernung in %	10,4	10,3	15,0	3,6
Durchschnittliche Entfernung in %	-1,8	-2,1	0,3	-3,8

Auswertung:

Der Unterschied in der maximalen Entfernung beträgt im Durchschnitt 10 %. Für den dargestellten Bereich sind somit bei der Knotenplatzierung im Mittel zwei Zentren mehr notwendig, um die Flexibilität der freien Platzierung zu kompensieren. Hierbei ist jedoch zu beachten, dass weitere Zentren auch zu einer Erhöhung der Kapazität führen und dem Lastausgleich dienen. Bei der Optimierung hinsichtlich der durchschnittlichen Entfernung erreicht der Algorithmus Dragoon mit der Anpassung auf die Nebenbedingung Knotenplatzierung im Durchschnitt eine 1,8 % bessere Leistung als die Anpassung für eine freie Platzierung. Mit steigender Anzahl von Zentren im Vergleich zu den Knoten tritt ein Sättigungseffekt ein, wodurch die absoluten Unterschiede vernachlässigbar sind. Das gleiche Verhalten von Knotenplatzierung und freier Platzierung zeigt sich auch bei der Nutzung des Algorithmus Lloyd, siehe Anhang E.1. Demgegenüber erreicht der Algorithmus MacQueen eine Verbesserung bei mehr Flexibilität, siehe Anhang E.2. Dies verdeutlicht die Komplexität der Problemstellung und zeigt weiteres Potential bei den Optimierungsalgorithmen. Darüber hinaus wird dadurch deutlich, dass im Allgemeinen ein an das Optimierungskriterium angepasster Algorithmus nicht notwendigerweise eine bessere Leistung erbringt. Auf Grund dem gleichen Stagnierungsverhalten bei der

knotenbasierten Platzierung als auch bei der freien Platzierung erfolgen die weiteren Analysen vorrangig anhand der Knotenplatzierung.

6.5.6 Vergleich der Auswirkung des Optimierungskriteriums bei der Platzierung

Basierend auf der Gesamtheit aller Testszenarien erfolgen detailliertere Analysen des entwickelten Algorithmus Dragoon in Bezug auf das Verhalten bei unterschiedlichen Optimierungskriterien. Die Abbildungen 6.15 und 6.16 zeigen die Leistungsdifferenz für die maximale und durchschnittliche Entfernung bei abweichendem Zielkriterium. Hierfür wird die Leistung des Algorithmus bei den Nebenbedingungen Knotenplatzierung und freie Platzierung gegenübergestellt. Dies erfolgt anhand den Optimierungskriterien maximale Entfernung (MAX) und durchschnittlichen Entfernung (AVG).

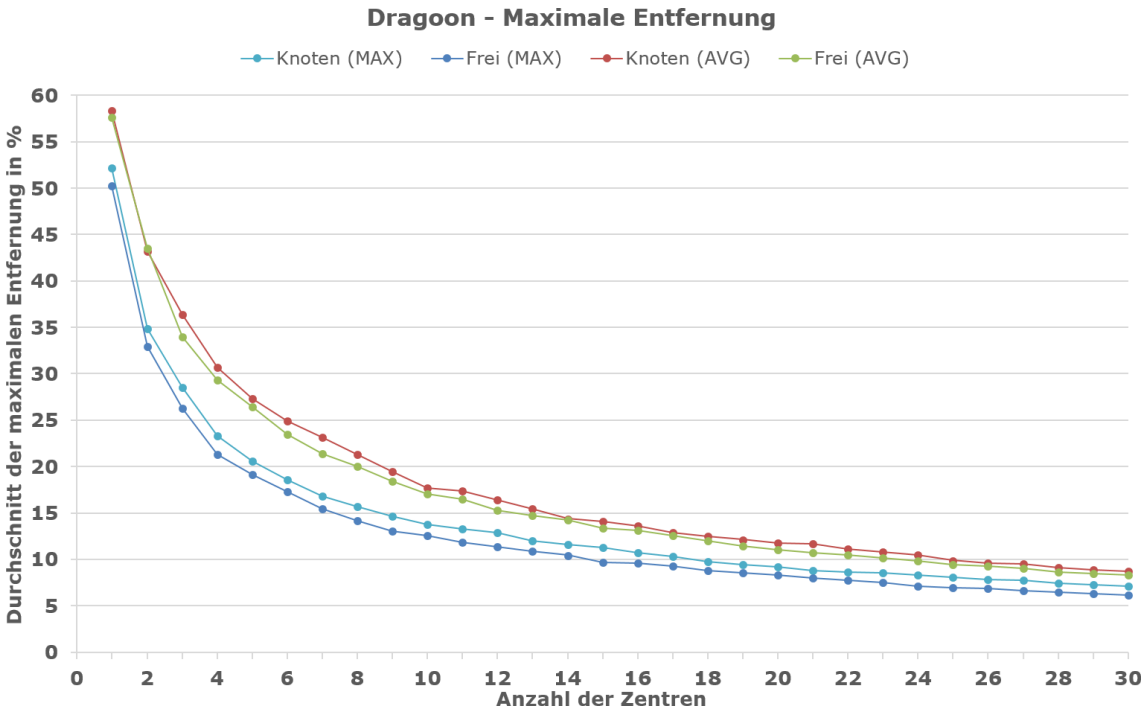


Abbildung 6.15: Einfluss unterschiedlicher Zielkriterien bei der Optimierung auf die maximale Entfernung bei den Nebenbedingungen knotenbasierte und freie Platzierung (Kleinere Werte sind besser.)

Auswertung:
Wie auch bei der vorherigen Analyse im Abschnitt 6.5.5 zu sehen ist, gibt es einen Leistungsunterschied in Abhängigkeit der Platzierungsbedingung. In Anbetracht der absoluten Werte zeigt sich, dass der Unterschied beim Optimierungskriterium Durchschnitt vernachlässigbar ist, sodass eine Optimierung hinsichtlich des Kriteriums der maximalen Entfernung eher von Bedeutung ist. Dies ist jedoch stets im Kontext des Anwendungsgebietes und der spezifischen Zielstellung für ein Szenario zu betrachten.

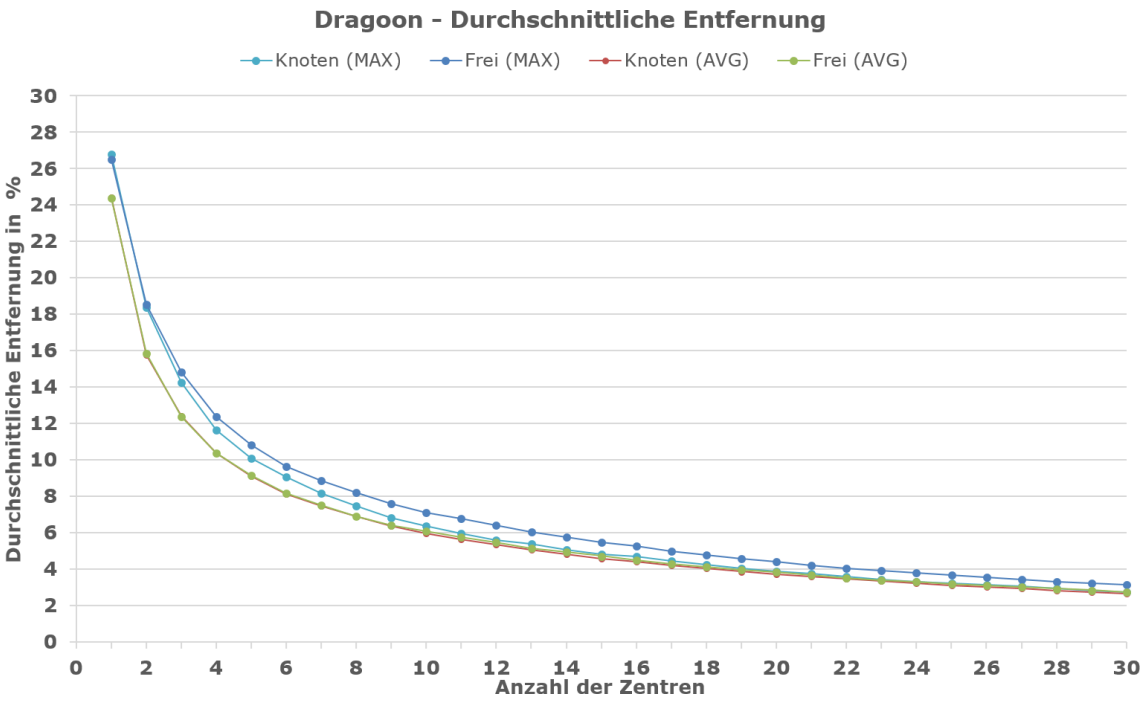


Abbildung 6.16: Einfluss unterschiedlicher Zielkriterien bei der Optimierung auf die durchschnittliche Entfernung bei den Nebenbedingungen knotenbasierte und freie Platzierung (Kleinere Werte sind besser.)

6.5.7 Analyse der Berechnungszeit

Die folgende Analyse vergleicht den praktischen Berechnungsaufwand der verschiedenen Algorithmen. Hierbei sind zwei verschiedene Einflussfaktoren zu unterscheiden, welche einen maßgeblichen Einfluss auf die Komplexität des Problems und damit auf die Berechnungszeit haben. Zum einen besteht eine Abhängigkeit der Berechnungszeit von der Anzahl der Zentren, welche zu platzieren sind. Zum anderen hängt die Komplexität von den Positionen und der Anzahl der Nutzer eines Szenarios ab. Die theoretische Berechnungskomplexität ist bei der Beschreibung der jeweiligen Algorithmen in Abschnitt 4.8 und 5.3 mit angegeben.

Die Bewertung des Einflusses von der Anzahl der zu planenden Zentren auf die Berechnungszeit erfolgt anhand des Szenarios Dialtelecom mit euklidischer Entfernungsfunktion und dem Optimierungsziel der Reduktion der maximalen Entfernung. Die Abbildungen 6.17 und 6.18 stellen die Berechnungszeit verschiedener Algorithmen vergleichend gegenüber. Auf Grund der Übersichtlichkeit und der unterschiedlichen Zeitskala ist dies in zwei Abbildungen aufgeteilt. Die Berechnungszeit in Abhängigkeit der Anzahl von Nutzern und deren Position wird mit den synthetischen Szenarien *Gephi 1 - 4* durchgeführt. Die Abbildungen 6.19 und 6.20 zeigen dazu den Vergleich für die verschiedenen Algorithmen. Die Ermittlung der durchschnittlichen Berechnungszeit für die Optimierung einer Aufgabenstellung erfolgt anhand von 100 Wiederholungen. Ausnahme bildet hier das Verfahren ILP mit nur 10 Wiederholungen auf Grund des hohen Zeitbedarfs. Für das Experiment ist das Rechnersystem entsprechend konfiguriert, sodass möglichst wenig Störeinflüsse während der Zeitmessung auftreten und die Rechenleistung

konstant bleibt. Nichtsdestotrotz sind diese nicht auszuschließen. Zu den verbleibenden Störeinflüssen und auftretenden Messschwankungen gehören unter anderem die *Garbage Collection* von Java zur automatischen Speicherbereinigung sowie Threadwechsel durch die CPU und die Einteilung der Zeitskala. Zudem ist die Art der Implementierung von den Algorithmen für die praktische Berechnungszeit ausschlaggebend.

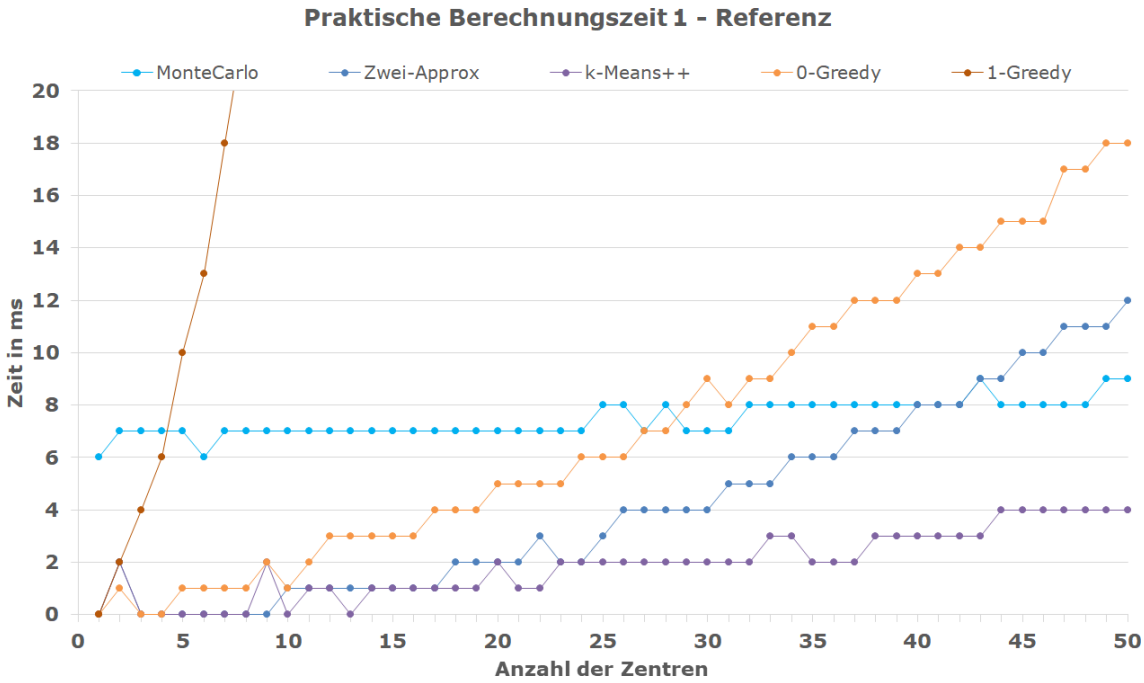


Abbildung 6.17: Durchschnittlicher Berechnungsaufwand für eine Ausführung des jeweiligen Algorithmus von den Referenzalgorithmen (Kleinere Werte sind besser.)

Auswertung (Anzahl der zu planenden Zentren):
Als Referenz für die Berechnungszeit dient die Monte Carlo Methode. Diese besitzt einen nahezu konstanten Aufwand für die unterschiedlichen Anzahlen zu platzierender Zentren. Beim Zwei-Approx Algorithmus als auch bei den Verfahren der k-Means Gruppe steigt der Berechnungsaufwand geringfügig mit der Anzahl der Zentren an. Dies trifft auch für 0-Greedy zu, wobei dieser Ansatz von den zuvor genannten proportional am stärksten steigt. Weiterhin wird für die δ-Greedy Verfahren deutlich, dass der Einfluss von Backtracking mit steigendem k die Berechnungszeit exponentiell steigen lässt. Bereits bei δ=1 und geringer Anzahl an Zentren ist der Aufwand gegenüber den anderen Algorithmen sichtlich erhöht. Der initiale Berechnungsaufwand für den entwickelten Algorithmus Dragoon ist bei der knotenbasierten Platzierung im Vergleich zu 0-Greedy und den k-Means Verfahren erhöht. Dies liegt an den mehreren Berechnungsphasen, der spezifischeren Vorgehensweise und der Evaluation des gesamten Szenarios in jeder Iteration, wobei die Implementierung nicht auf geringe Berechnungszeit vollends optimiert ist. In Abhängigkeit der Komplexität sind bei verschiedenen Anzahlen von Zentren mal mehr und mal weniger Iterationen notwendig, was sich in der schwankenden Berechnungszeit bemerkbar macht. Mit steigender Anzahl von Zentren erhöht sich der Berechnungsaufwand, wobei dieser geringer Ausfällt als bei 1-Greedy. Bei der freien Platzierung mittels Dragoon ist der Berechnungsaufwand deutlich höher auf

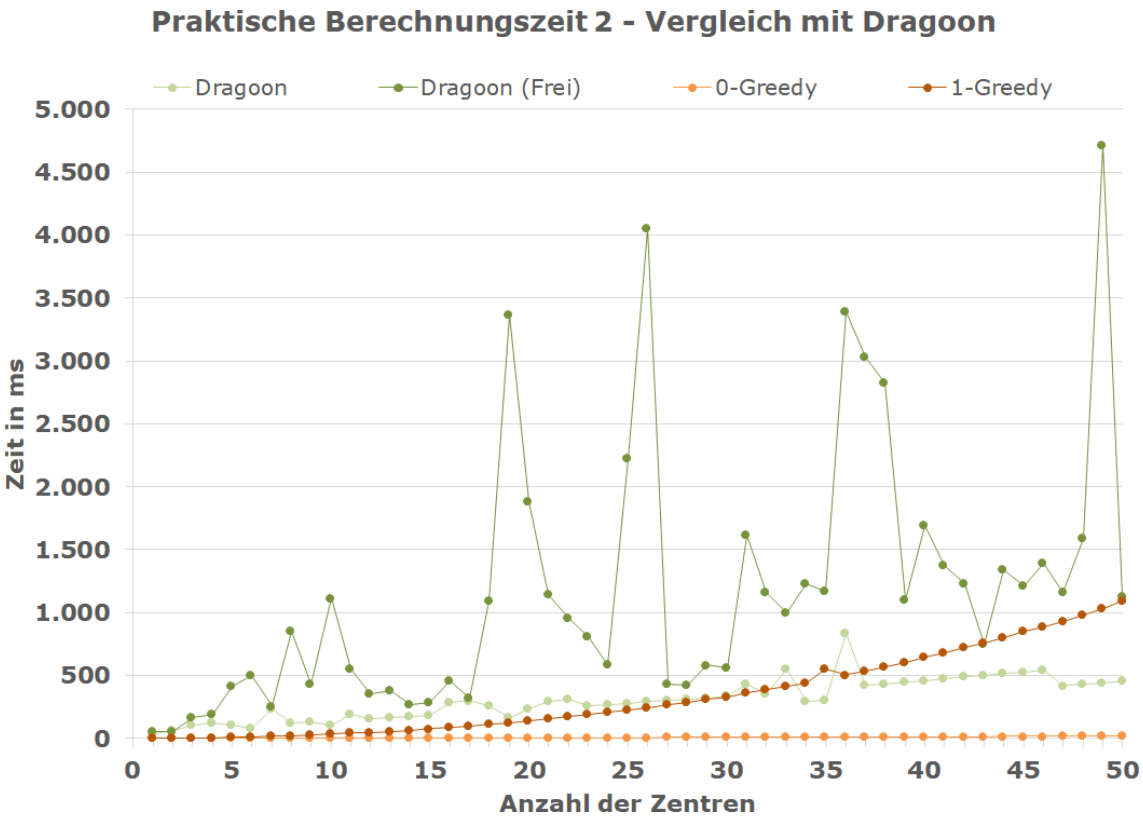


Abbildung 6.18: Durchschnittlicher Berechnungsaufwand für eine Ausführung des jeweiligen Algorithmus gegenüber den verschiedenen Varianten von Dragoon (Kleinere Werte sind besser.)

Grund der Diskretisierung des Lösungsraumes. Die Berechnung wurde hierbei mit der Genauigkeit 10^{-7} durchgeführt, wobei dies einen maßgeblichen Einfluss auf die Zeitbedarf hat. Hierbei wird die notwendige Anzahl an Iterationen bis zur endgültigen Stagnation in Abhängigkeit der Problemstellung ersichtlich. Der wesentliche Unterschied zwischen Dragoon und den anderen heuristischen Verfahren besteht in dem Einfluss von Zufallsentscheidungen. Dragoon liefert bereits bei der ersten Ausführung ein stabiles Ergebnis, wohingegen die anderen Verfahren für ein aussagekräftiges Ergebnis mehrfach auszuführen sind.

Die Laufzeit der evolutionären Algorithmen hängt maßgeblich von den zugestandenen Generationen und den spezifizierten Populationsgrößen ab. Durch Vorabtests erfolgte eine Ermittlung der an die Problemstellung angepassten Parameter für den GA als auch SA. Die Berechnungszeiten für 2000 Iterationen liegt beim GA und SA zwischen 8 und 12 Sekunden.

Die höchsten Berechnungszeiten in diesem Experiment hat der deterministische Ansatz über das ILP Verfahren. Diese liegen bei über einer Minute bei bereits geringen Anzahlen von Zentren und steigen überproportional mit jedem weiteren Zentrum. Auf Grund der Übersichtlichkeit sind die hohen Werte von den evolutionären Algorithmen als auch vom ILP nicht mit in den Abbildungen eingezeichnet.

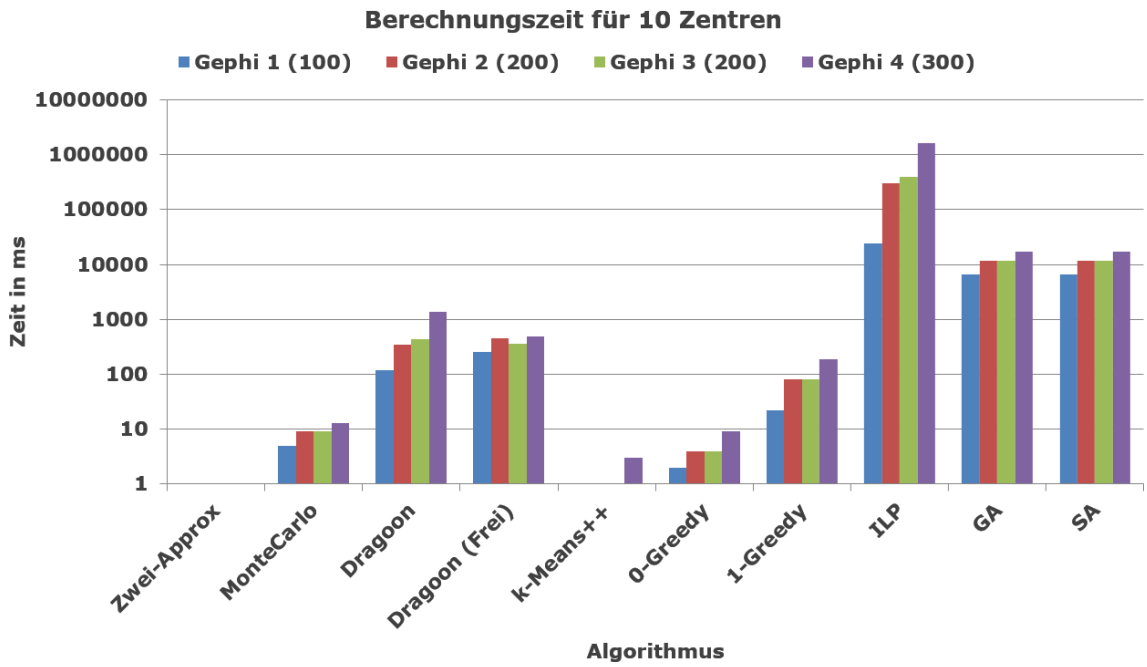


Abbildung 6.19: Durchschnittlicher Berechnungsaufwand für eine Ausführung des jeweiligen Algorithmus bei unterschiedlicher Anordnung und Anzahl von Nutzern für 10 Zentren (Kleinere Werte sind besser.)

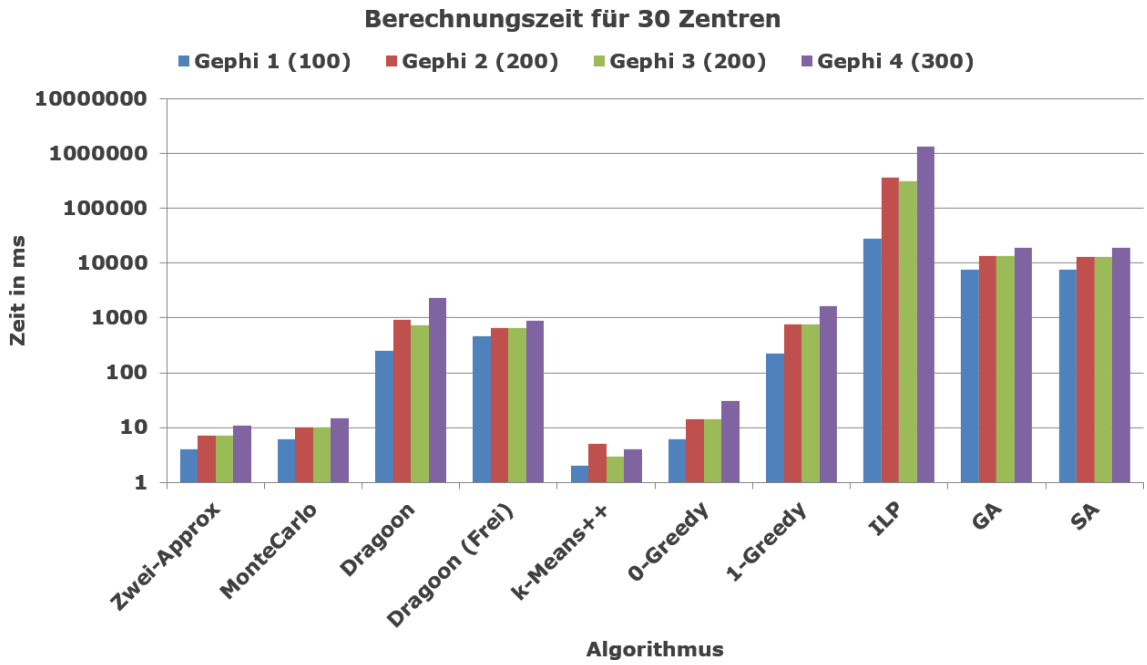


Abbildung 6.20: Durchschnittlicher Berechnungsaufwand für eine Ausführung des jeweiligen Algorithmus bei unterschiedlicher Anordnung und Anzahl von Nutzern für 30 Zentren (Kleinere Werte sind besser.)

Auswertung (Anzahl und Position der Nutzer):
Unabhängig vom Algorithmus hängt die Berechnungszeit von der Komplexität des Szenarios ab. Mit steigender Anzahl von Nutzern steigt auch der Berechnungsaufwand. Dies trifft für alle hier betrachteten Optimierungsverfahren zu.

In Bezug auf die Anordnung der Nutzer mit deren Standort ergeben sich bei gleicher Anzahl für einige der Algorithmen unterschiedliche Berechnungszeiten. Dies ist anhand der Szenarien Gephi 2 und 3 mit jeweils 200 Nutzer ersichtlich. Die Algorithmen lassen sich diesbezüglich in zwei Gruppen einteilen. Bei der ersten Gruppe mit Zwei-Approx, Monte Carlo, Greedy und Evolutionären Algorithmen ist die Laufzeit unabhängig von der Position der einzelnen Nutzer. Die zweite Gruppe mit Dragoon, k-Means und ILP reagiert auf die spezifische Situation, wobei unterschiedliche Anzahlen an Iterationen zur vollständigen Optimierung notwendig sind. Somit beeinflusst die Anordnung die Stagnation der Optimierungsalgorithmen. Dies spiegelt sich in der unterschiedlichen Berechnungszeit wider.

6.5.8 Allgemeiner Leistungsvergleich

Zum allgemeinen Leistungsvergleich zeigt die Tabelle 6.8 die Entfernung bei den vereinigten Testdaten von den verschiedenen Optimierungsverfahren. Die Analyse erfolgt anhand der knotenbasierten Platzierung, da hierfür mehr Algorithmen zum Vergleich zur Verfügung stehen. Im Anhang E.3 sind die minimalen und maximalen Werte von den mehreren Wiederholungen aufgeführt. Die Leistungsunterschiede bei verschiedenen Nebenbedingungen sind in Abschnitt 6.5.5 aufgezeigt. Gemäß der Analyse in Abschnitt 6.5.4 wird beim Algorithmus Dragoon+ gegenüber der normalen Variante die Reihenfolge nach jeder Iteration umgekehrt, in der die Zentren betrachtet werden.

Auswertung:
Die Lösungsansätze mittels Monte Carlo und Zwei-Approx stellen mit deren Ergebnissen die Referenz dar. Insbesondere zeigt sich beim Zwei-Approx auf Grund der Vorgehensweise eine stetige Verbesserung der Leistung mit steigender Anzahl an Zentren. Beachtenswert ist die Leistung der evolutionären Algorithmen des Frameworks SEREIN. Diese sind nicht spezifisch für die Problemstellungen konzipiert und erreicht dennoch vergleichbare Lösungen gegenüber anderen Algorithmen, welche spezifisch für diese Probleme entwickelt wurden. Die Leistung der Cluster-Verfahren MacQueen, Llyod und K-Means++ ist ähnlich, wobei K-Means++ auf Grund der angepassten Initialisierung im Schnitt die besseren Lösungen von den drei Verfahren erreicht. Weiterhin zeigt sich bei dem Ansatz von ILP, dass dieser trotz der garantierten Lösungsgüte nicht die besten Ergebnisse erreicht. Die Algorithmen Greedy und GRASP erreichen im Vergleich nur eine mäßige Leistung. Die besten Ergebnisse erzielen im Durchschnitt die Dragoon basierten Ansätze. Deutliche Unterschiede gibt es bei der Komplexität und damit in der Berechnungsdauer von den Algorithmen, wobei alle Optimierungen bei den Testdatensätzen innerhalb von mehreren Stunden beendet wurden.

Tabelle 6.8: Leistungsvergleich für das Zielkriterium Minimierung der maximalen Entfernung in % (Kleinere Werte sind besser.)

Anzahl der Zentren	Monte Carlo	Zwei-Approx	ILP	0-Greedy	GRASP	MacQueen	Lloyd	k-Means++ (Frei)	GA	SA	**Dragoon**	**Dragoon+**	**Dragoon (Frei)**	**Backtrack**
1	73,6	73,6	95,2	52,2	52,2	55,6	55,6	54,8	52,2	52,2	52,2	52,2	50,3	52,2
2	58,5	51,7	54,1	49,2	36,0	41,9	41,1	40,7	34,4	34,4	34,8	34,8	32,9	34,5
3	51,1	41,2	39,2	39,8	31,0	36,3	34,4	33,3	28,0	27,8	28,5	28,5	26,3	27,7
4	46,2	33,7	32,6	34,8	27,9	32,2	29,1	28,1	23,4	22,9	23,3	23,3	21,3	22,5
5	42,8	28,7	28,3	30,8	25,7	29,8	26,4	25,4	20,7	20,3	20,6	20,5	19,1	19,9
10	33,0	17,3	17,5	19,3	20,2	23,5	18,5	16,8	14,5	13,9	13,8	13,5	12,5	12,8
15	28,4	13,1	13,5	14,5	17,5	19,9	14,9	13,0	11,7	11,2	11,3	10,8	9,7	10,0
20	24,9	10,6	11,0	11,8	16,0	17,6	12,9	10,8	10,0	9,4	9,2	8,9	8,3	8,3
25	22,6	8,9	9,2	10,1	14,9	16,2	11,4	9,2	8,8	8,4	8,0	7,6	7,0	7,1
30	20,7	7,7	8,0	8,7	14,0	15,2	10,2	8,0	8,0	7,5	7,1	6,7	6,2	6,2
35	19,3	6,8	7,1	7,7	13,2	14,6	9,3	7,1	7,3	6,8	6,4	5,9	5,4	5,6
40	17,9	6,1	6,3	6,9	12,7	13,9	8,6	6,4	6,7	6,2	5,8	5,4	4,9	5,0
45	17,0	5,5	5,7	6,2	12,0	13,3	8,0	5,6	6,2	5,9	5,2	4,9	4,4	4,6
50	16,0	4,9	5,1	5,6	11,3	12,7	7,5	5,1	5,8	5,5	4,7	4,5	3,8	4,2

6.5.9 Analyse der notwendigen Anzahl von Zentren hinsichtlich der Entfernung

Mit steigender Anzahl von Zentren tritt ein Sättigungseffekt hinsichtlich der Reduktion von Entfernungen ein. Daher erfolgt eine Analyse für die effektive Anzahl von Zentren. Die Reduktion der Entfernung pro zusätzlichem Zentrum k wird gemäß der Gleichung 6.7 berechnet.

$$\Delta D = (D_k - D_{k+1})/D_k \cdot 100 \tag{6.7}$$

Die Tabelle 6.9 und die Abbildung 6.21 zeigen die akkumulierten Ergebnisse bei den synthetischen Szenarien mit 200 Nutzern (Index 3, 4, 6, 7). Die Platzierung erfolgt knotenbasiert.

Tabelle 6.9: Verbesserung der Entfernung in Relation zur Anzahl der Zentren mittels Dragoon anhand der Knotenplatzierung (Rundung der Werte gemäß der Originaldaten)

Anzahl Zentren	1	2	3	4	5	10	15	20	25	30	35	40	45	50
Maximale Entf. in %	52,0	39,0	33,7	25,5	24,4	15,9	13,2	11,0	10,1	9,4	8,6	8,1	7,5	7,2
Verbesserung in %	-	24,8	13,6	24,3	4,6	7,0	3,4	3,4	1,5	1,5	1,6	1,2	1,5	0,8
Durchschn. Entf. in %	28,5	21,1	17,8	14,7	13,2	8,5	6,9	5,6	4,8	4,3	3,8	3,4	3,2	2,8
Verbesserung in %	-	26,2	15,4	17,2	10,2	7,1	3,9	3,8	2,5	2,4	2,3	2,3	1,6	2,0

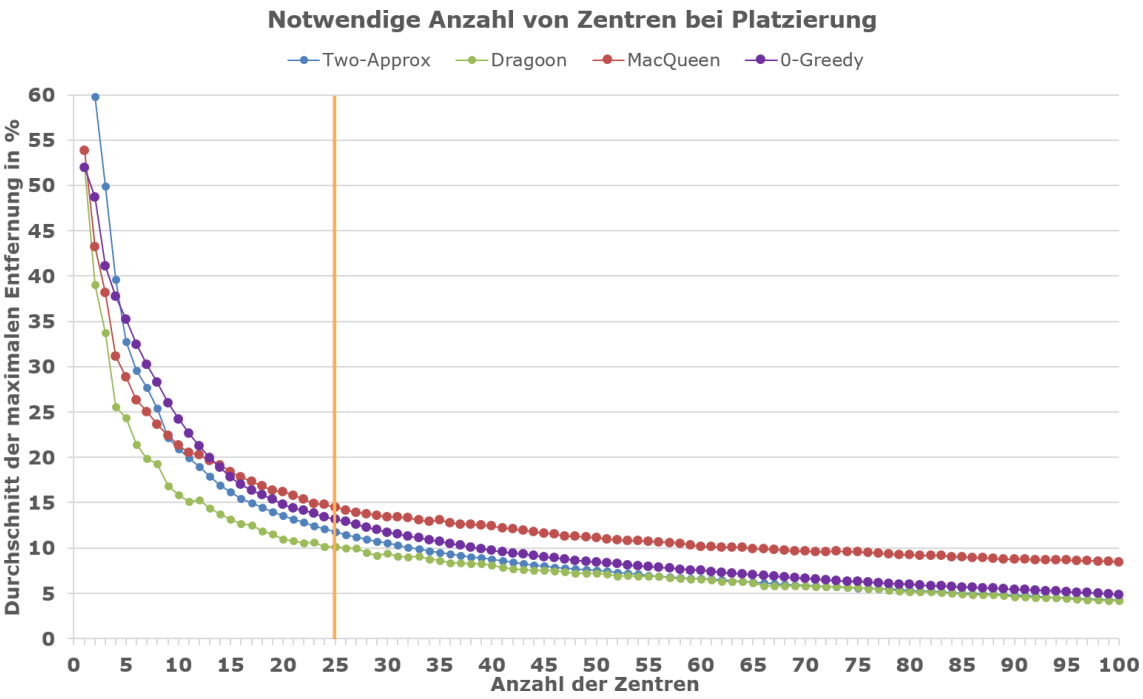


Abbildung 6.21: Leistungsübersicht hinsichtlich der maximalen Entfernung: Die orange, vertikale Linie markiert die Anzahl von 25 Zentren, wobei die durchschnittliche Verbesserung danach weniger als 2 % pro zusätzlichem Zentrum beträgt. (Kleinere Werte sind besser.)

Auswertung:
Die Analyse zeigt, dass es ausreichend ist, bei weniger als ca. 15 % der gesamten Menge von Nutzern ein Zentrum einzurichten. Bereits bei 25 Zentren erreicht der Algorithmus Dragoon bei diesen normalisierten Testdaten mit jedem weiteren Zentrum nur noch eine Verbesserung der maximalen Entfernung eines Nutzers von unter 2 %. Nachdem 49 Zentren platziert wurden, beträgt die durchschnittliche Verringerung der Entfernungen bei einem weiteren Zentrum im Schnitt weniger als 1 %. Bei Zwei-Approx sind trotz der garantierten Approximationsgüte auf Grund der schlechteren Platzierungsleistung 73 Zentren notwendig, bis eine Sättigung eintritt. Somit wird deutlich, dass die Leistung des verwendeten Algorithmus Einfluss auf die notwendige Anzahl von Zentren bei einem Szenario hat. Allgemein verringert sich die Anzahl der Zentren mit besserer Leistung des Optimierungsalgorithmus.

6.5.10 Analyse zur operationellen Planung

In Anbetracht der Anwendungsgebiete und der initialen Planung von Zentren, sind die operativen Kosten des Gesamtsystems zu berücksichtigen. Dafür ermöglicht das generische Modell eine Analyse, welche die Transportkosten auf Grundlage der Entfernungen der Nutzer zu den Zentren als auch die Betriebskosten für die Zentren in Relation betrachtet. Mit steigender Anzahl von Zentren verringert sich im Allgemeinen die durchschnittliche Entfernung für den Transfer und damit die Transportkosten und -zeit, wohingegen die Einrichtungs- und Betriebskosten steigen. Um die optimale Balance zwischen diesen Aspekten zu finden, erfolgt eine simulationsbasierte Optimierung zur

Berechnung der optimalen Anzahl von Zentren mit den dazugehörigen Standorten. Die Abbildung 6.22 zeigt für das Szenario Schweiz aus dem Testdatensatz der Logistik die Analyse. Hierbei wurde eine abstrakte Kostenfunktion angesetzt, welche in Abhängigkeit der Eingabedaten flexibel austauschbar ist, um die Möglichkeiten des generischen Modells zu verdeutlichen. Diese Berechnung hat auf Grund der individuellen Eigenschaften für jedes Szenario separat zu erfolgen, sodass hier nur ein Beispiel angegeben wird.

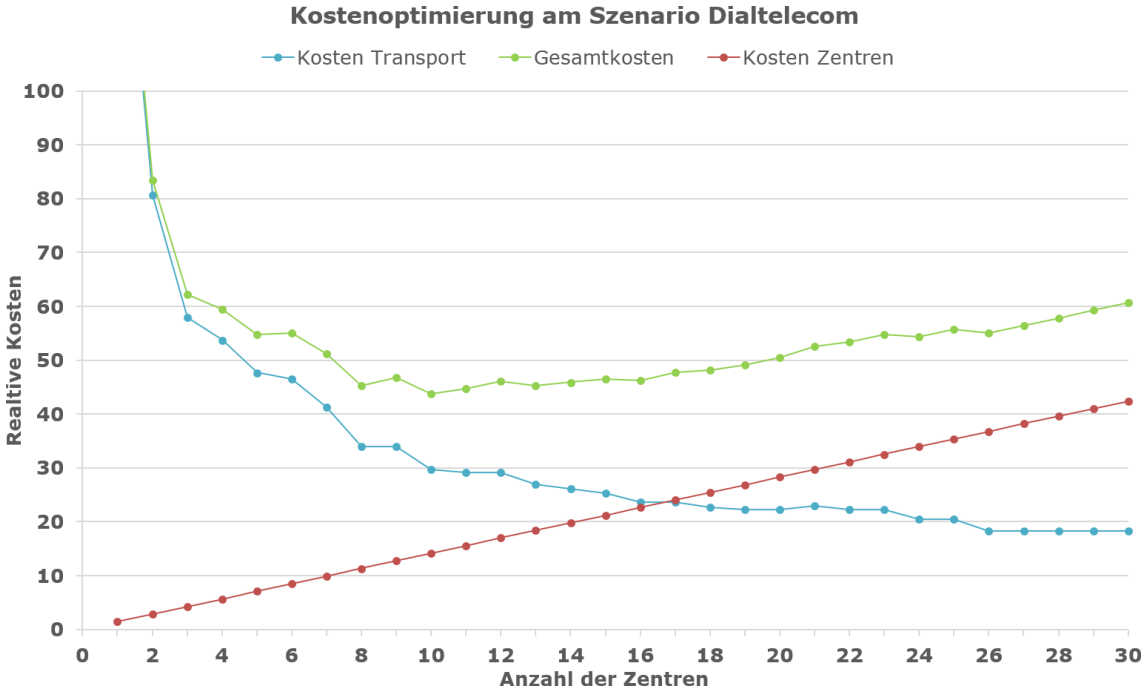


Abbildung 6.22: Kostenoptimierung am Szenario Schweiz (Kleinere Werte sind besser.)

Auswertung:
Durch die Akkumulation der Kosten und dem Wissen über die weiteren Leistungseigenschaften einer optimierten Lösung bei entsprechender Anzahl von Zentren lässt sich eine Entscheidung über die umzusetzende Lösung treffen. In dem Beispiel ergibt sich die günstigste Lösung bei der Planung von zehn Zentren. In Abhängigkeit der Leistung des Optimierungsalgorithmus ermöglicht eine solche Analyse eine Bewertung darüber, ob es auf Grund der nicht linearen Verbesserung lohnenswert ist, ein weiteres Zentrum zu planen.

6.6 Evaluierung der Geolokalisation

Auf Grundlage des spezifischen Modells der Geolokalisation wird dessen Funktionsweise evaluiert. Dazu erfolgt ein Vergleich mit bestehenden aktiven und passiven Verfahren. Insbesondere wird hierbei der Zusammenhang zum generischen Modell ersichtlich sowie dessen Flexibilität aufgezeigt. In mehreren Experimenten erfolgt die Geolokalisation von Zielen in Europa, da hierfür öffentliche Informationen zur Netzinfrastruktur zur Verfügung stehen [19]. Die Latenzmessungen werden über mehrere öffentliche Netze verschiedener ISPs durchgeführt, wodurch es sich um reale Anwendungsbeispiele unter realistischen

Bedingungen handelt. Dabei werden IP-Adressen von privaten Nutzern mit moderaten Internetverbindungen in die Zielbestimmung mit einbezogen. Die Messungen erfolgen mit Hilfe des RIPE Atlas Projektes [329], dass über 8200 VPs zur Verfügung stellt. Entsprechend der TIER 1 Topology erfolgt über den Algorithmus Dragoon die Auswahl der optimierten Messpunkte, welche dann direkt beziehungsweise auf die nächstgelegenen VP abgebildet sind. Anschließend wird entsprechend dem Modell der Geolokalisation im Abschnitt 5.2.3 vorgegangen. Um die Anzahl der Hops und die RTT zu bestimmen, werden Traceroute, Paris Traceroute und ICMP-Echo-Requests Pakete gemäß der Schnittstelle des RIPE Atlas Projektes versendet. Für die Experimente wird jede RTT Messung mit zehn Anfragen ausgeführt und jede Pfadlänge einmal mit jedem Verfahren bestimmt.

6.6.1 Einfluss der Entfernungsfunktion und Auswahl der Vantage Points

Die Tabelle 6.10 zeigt die Auswirkung unterschiedlicher Herangehensweisen auf die Genauigkeit bei der Standortbestimmung von mehreren Zielen. Dabei werden verschiedene Verfahren zur Auswahl von VPs sowie die zu Grunde gelegten Funktion zur Entfernungsberechnung auf die IP Geolokalisation vergleichend gegenüber gestellt. Das verwendete Modell der Geolokalisation bleibt hierbei gleich. Die Auswahl der VPs erfolgt gemäß den Algorithmen Dragoon, ILP und Zwei-Approx. Zur Entfernungsberechnung kommen der Kartendienst Google Maps (GM) sowie die Funktion für Orthodrome (OD) zum Einsatz. Zusätzlich erfolgt eine Überprüfung des spezifischen Modells der Geolokalisation mittels dem spezifischen Ansatz gemäß dem hyperbolischen Verfahren mit minimaler quadratischer Abschätzung (HYP) [300]. Dafür werden die ermittelten VPs gemäß Dragoon verwendet. Zudem zeigt ein Vergleich der Abbildungsfunktionen mit der Entfernungsapproximation gemäß dem Straßennetz beziehungsweise entlang der Erdoberfläche den Einfluss auf die Genauigkeit, siehe Anhang F.2. Weitere Informationen zu den Zielen sind im Anhang F.3 aufgeführt.

Auswertung:
Im Bezug zur Standortwahl der VPs zeigt sich, dass eine optimierte Platzierung einen deutlichen Einfluss auf die Messergebnisse und somit auf die Genauigkeit der Standortbestimmung hat. Die Platzierung der VPs mittels dem entwickelten Algorithmus Dragoon führt zu besseren Ergebnissen als bei Verwendung von VPs nach den Verfahren ILP und Zwei-Approx. Zudem wird der Einfluss der verwendeten Entfernungsfunktion auf die Genauigkeit bei gleichen VPs ersichtlich. Die Approximation der Verbindungslänge im Netz über das Straßennetz führt im Mittel zu einer präziseren Standortbestimmung als bei Nutzung von orthodromen Entfernungen. Dies zeigen insbesondere die genaueren Fälle mit den VPs gemäß Dragoon und ILP. Zudem ist dieser Effekt auch bei Verwendung des spezifischen Ansatzes nach HYP ersichtlich. Eine Ausnahme bilden die Messungen von Zwei-Approx auf Grund der ungünstigen Platzierung und dem damit verbundenen Einfluss der „Letzten Meile“ mit den ohnehin ungenaueren Ergebnissen.

Weiterhin ist dieser Unterschied bei der Abbildung von Zeit zu einer Entfernung in der Krümmung der Kurve ersichtlich. Gemäß dem physikalischen Zusammenhang folgt dieser bei Berücksichtigung aller Nebeneffekte im idealen Fall einer linearen Kurve. Die Abbildungsfunktion mit orthodromen Entfernungen ist stärker gekrümmt als die

Tabelle 6.10: Vergleich der Genauigkeit des spezifischen Modells zur IP Geolokalisation mit 20 VPs bei Verwendung unterschiedlicher Landmark Positionen und Approximationen für die Entfernung (Alle Angaben der Abweichungen erfolgen in km. Kleinere Werte sind besser.)

Ziel	Dragoon (GM)	Dragoon (OD)	ILP (GM)	ILP (OD)	Zwei-Approx (GM)	Zwei-Approx (OD)	HYP(GM)	HYP(OD)
Amsterdam	59	4	67	799	60	60	769	840
Ballerup	48	55	109	38	71	121	339	414
Berlin	244	288	215	312	36	175	438	570
Calw	56	221	242	205	537	229	636	938
Delft	65	65	144	632	667	799	730	815
Ettlingen	126	462	225	210	318	153	743	313
Frankfurt	45	153	588	803	94	93	236	803
Glattbrugg	34	7	6	6	275	159	228	494
Göteborg	328	120	370	333	324	470	295	333
Hamburg	16	245	226	197	266	131	361	340
Kiel	109	242	129	476	254	294	728	388
Prag	132	152	217	218	383	365	595	210
Stuttgart	322	254	718	804	666	547	247	895
Turin	23	268	17	223	851	544	381	674
Wien	0	0	251	180	177	82	605	240
Zürich	7	1	260	297	277	146	287	306
∅ Abw.	101	159	237	358	329	273	476	536

Kurve gemäß Google Maps. Dies verdeutlicht zudem den Einfluss der „Letzten Meile“, was im begrenzten Maße durch die Entfernungsapproximation mittels Straßennetz zu einer genaueren Modellierung führt. Daher verwendet das spezifische Modell der Geolokalisation bei allen weiteren Standortbestimmungen das Straßennetz als Grundlage der Entfernungsberechnung.

6.6.2 Vergleich der Genauigkeit mit anderen Diensten der IP Geolokalisation

Die Tabelle 6.11 zeigt den Vergleich mit anderen Diensten. Hierbei werden die passiven Dienste WHOIS [286] und MaxMind [291] mit den aktiven Techniken Spotter [280] und TULIP [330] der Stanford University (CBG, TBG, Geoplugin und FreeGeoIP) verglichen.

Auswertung:
Die Messergebnisse zeigen die Präzision des vorgestellten Modells zur Geolokalisation als Gesamtkonzept in Bezug zum Stand der Technik und weiteren Evaluierungen [281, 331]. Die beiden passiven Ansätze erreichen im Vergleich zum vorgestellten aktiven Ansatz ebenfalls überwiegend hohe Genauigkeiten. Die anderen aktiven Verfahren zeigen demgegenüber im Durchschnitt größere Abweichungen. Da die aktiven Referenzverfahren nicht direkt zur Verfügung standen, erfolgte die Bestimmung mittels dem Online-Dienst

Tabelle 6.11: Vergleich der Genauigkeit des spezifischen Modells zur IP Geolokalisation mit 20 VPs gegenüber anderen aktiven (A) und passiven (P) Diensten (Alle Angaben der Abweichungen erfolgen in km. Kleinere Werte sind besser.)

Ziel	*Modell gemäß Abschnitt 5.2.3 (A)*	*Whois (P)*	*MaxMind (P)*	*Spotter (A)*	*CBG (A)*	*TBG (A)*	*Geoplugin (A)*	*FreeGeoIP (A)*
Amsterdam	59	*5481**	*5474**	477	172	163	5390*	5389*
Ballerup	48	*14*	*13*	738	806	795	13	22
Berlin	56	*2*	*341*	401	866	866	347	347
Calw	244	*1*	*13*	46	46	46	13	6
Delft	65	*16*	*65*	495	205	205	65	65
Ettlingen	126	*0*	*1*	111	4	4	1	1
Frankfurt	45	*379*	*370*	136	101	101	366	366
Glattbrugg	34	*1*	*64*	221	109	109	64	64
Göteborg	328	*69*	*402*	951	1063	1015	396	396
Hamburg	16	*24*	*3*	475	586	586	3	3
Kiel	109	*453*	*381*	567	676	771	382	382
Prag	132	*12*	*2*	335	459	459	2	2
Stuttgart	322	*28*	*254*	11	64	64	6	254
Turin	23	*433*	*477*	411	483	483	485	485
Wien	0	*14*	*1*	478	481	481	1	1
Zürich	7	*24*	*2*	193	179	179	2	2
∅ Abw.	101	*434*	*492*	378	394	395	471	487
∅ Abw.*	104	*98*	*159*	371	408	411	143	160

TULIP [330]. Dieser bietet für die verschiedenen Methoden Re-Implementierungen an. Die dabei verwendeten VPs sind entsprechend vom Anbieter selbst spezifiziert.

Bei den Ergebnissen fallen insbesondere die Abweichungen von über 5000 km auf. Grund dafür sind bei passiven Verfahren veraltete Informationen und unzureichende Datenbankeinträge, beziehungsweise ergeben Anfragen an diese Dienste keine Angaben zu einem Ziel. Dies ist vor allem durch die dynamische Zuweisung von IP-Adressen bedingt. Ein solcher Effekt tritt auch bei aktiven Verfahren durch fehlerhafte oder unzureichende Messungen auf. Diese entstehen durch Paketverluste in der Kommunikation, unterdrückten Antworten und der Filterung von dedizierten Protokollen mittels Firewalls und Proxys, wodurch es zu Zeitüberschreitungen kommt. In diesen Fällen wird als Ergebnis ein Standardwert mit den Geokoordinaten (Latitude=0, Longitude=0) zurück gegeben. Die durchschnittliche Abweichung, ohne die Ergebnisse vom Ziel Amsterdam, wird in der Zeile „ ∅ Abw.* “ angegeben. Auch dabei ergibt sich für das in Abschnitt 5.2.3 vorgestellte Modell eine bessere Genauigkeit als bei den anderen aktiven Verfahren. Der vorgestellte Ansatz beachtet im Vergleich zu den anderen Verfahren die Fluktuation der Netzverzögerungen durch parallelen Datenverkehr in öffentlichen Netzen und nutzt eine genauere Approximation der Entfernungen auf Grundlage des Straßennetzes.

Zudem erfolgt eine optimierte Auswahl der VPs, wobei die Anzahl der Hops in die Standortbestimmung mit einfließen.

6.7 Evaluierung des Gesamtsystems

Auf Grundlage des Konzeptes des generischen Modells erfolgt die Analyse und Bewertung hinsichtlich eines optimierten Gesamtsystems. Dazu werden verschiedene Tests durchgeführt, um allgemeine Effekte zu demonstrieren und aufzuzeigen. Insbesondere liegt hierbei der Fokus auf dem Leistungsunterschied zwischen optimierter Platzierung der Zentren ungeachtet des Nutzerverhaltens und optimierter Zuweisung der Nutzeranfragen anhand der Profile unabhängig des Standortes.

6.7.1 Analyse der Speicherverwaltungsstrategien in Abhängigkeit der Größe des Speichers

Bevor eine Evaluierung verschiedener Zusammenhänge hinsichtlich des operativen Nutzens des Gesamtsystems erfolgen, werden die Speicherverwaltungsstrategien des Speichers aus Abschnitt 5.1.12 bewertet. In dem Experiment wird die Größe des Zwischenspeichers in Abhängigkeit der Verwaltungsstrategie analysiert, um eine effiziente Funktionsweise zu erhalten. Da die Ergebnisse direkt von dem jeweiligen Szenario abhängen erfolgt keine Mittelwertbildung über mehrere Szenarien, um somit verschiedene Effekte aufzuzeigen. Die maßgeblichen Parameter der Simulation sind in der Tabelle 6.12 aufgeführt.

Tabelle 6.12: Parameter des Experiments Speicherersetzungsstrategien

Parameter	Variable
Szenario	Index 23 - Dialtelecom (Cz)
Anzahl der Zentren	5
Platzierungsstrategie	Dragoon
Entfernungsfunktion	Dijkstra mit Metrik Hops
Zielkriterium	Reduktion der maximalen Entfernung
Dienste im Universum	100
Anteil der Nutzerprofile	5 % und 15 % des Universums
Zipf-Verteilung	$\alpha = 0{,}75$
Anfragen jedes Nutzers	100

Die Zentren befinden sich stets am gleichen, optimierten Standort und die Nutzer sind jeweils dem nächstgelegenen Zentrum zugewiesen. Jedes Zentrum besitzt dabei die gleiche Speicherkapazität. Der Fokus liegt hierbei auf den Speicherersetzungsstrategien, da das Szenario ein CDN nachbildet und die Auswirkungen sich besser analysieren lassen als mittels Bestandsoptimierung. Die Simulation wird hierbei gemäß dem rekursiven Systemverhalten von Abschnitt 5.1.12 durchgeführt. Die Abbildungen 6.23 und 6.24 zeigen für verschiedene Ersetzungsstrategien die Zugriffsfehlerrate von den Nutzern beim Zugriff

auf das primäre Zentrum. Das Verhältnis aus allen Anfragen und den Antworten, wobei ein angeforderter Dienst nicht sofort im Speicher zur Verfügung steht, ergibt die Fehlerrate. Dabei wird die Leistung der Algorithmen bei verschiedenen Speichergrößen verglichen.

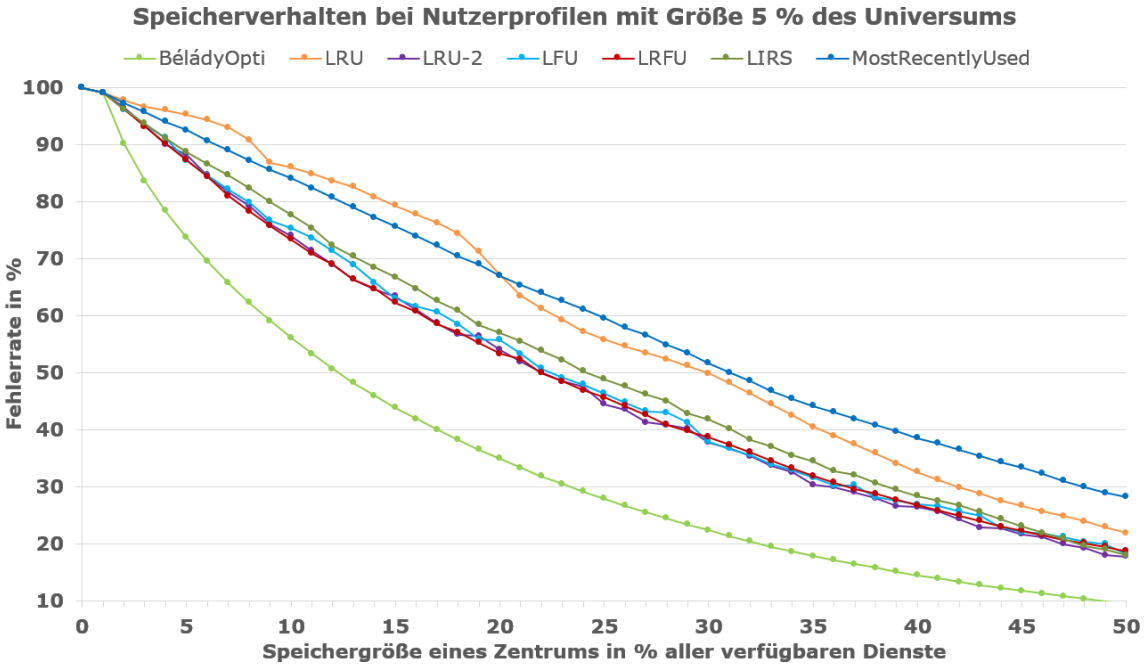


Abbildung 6.23: Leistungsvergleich von Speicherersetzungsstrategien gegenüber dem Optimum mittels Bélády bei verschiedenen Speichergrößen und 5 % Anteil der Nutzerprofile vom Universum

Auswertung:
Mit zunehmender Speicherkapazität ergibt sich auch eine Verbesserung der Fehlerrate. Die optimale Strategie ist vor allem bei geringer Speichergröße in der Lage diesen effizient zu nutzen, wohingegen der Verlauf der anderen Verfahren nahezu linear verläuft auf Grund der Diversität der Nutzerprofile. Bei der kleineren Größe der Profile mit 5 % des Universums verlaufen die Leistungskurven regressiver als bei 15 %. Mit verminderter Vielfältigkeit der Profile sind damit die Verwaltungsstrategien eher in der Lage für zukünftige Anfragen die richtigen Entscheidungen zu treffen. Durch die erhöhte Anzahl von Diensten im Profil werden auf Grund der Verteilung des Spektrums die anderen Dienste öfters angefragt.

Die Leistung der weiteren Algorithmen aus Abschnitt 5.1.12 liegt hierbei oberhalb von LRU-2 und sind auf Grund der Übersichtlichkeit nicht in der Abbildung eingetragen. Gemäß Bélády's Anomalie Phänomen zeigt sich, dass es bei Vergrößerung der Speicherkapazität nicht stets eine Verbesserung der Fehlerrate eintritt. Die Abbildung zeigt deutlich das Potential für Verbesserungen auf. Die Entwicklung der bisherigen Strategien zur Speicherverwaltung ist bisher nicht in der Lage die Leistungslücke zu schließen. Neben einer verbesserten Verwaltung und erhöhtem Einsatz von Ressourcen besteht die weitere Möglichkeit die Anfragen besser zu Koordinieren, mit dem Zweck die Anfragen nach gleichen Diensten an einem Zentrum zu erhöhen und somit die Fehlerrate zu verringern. Daher wird mit dem Ansatz in der vorliegenden Arbeit das Ziel verfolgt über die Korrelation der Nutzerinteressen sowie einer entsprechenden Anfragezuweisung eine Zusammenfassung gleicher Anfragen an den Zentren zu erreichen.

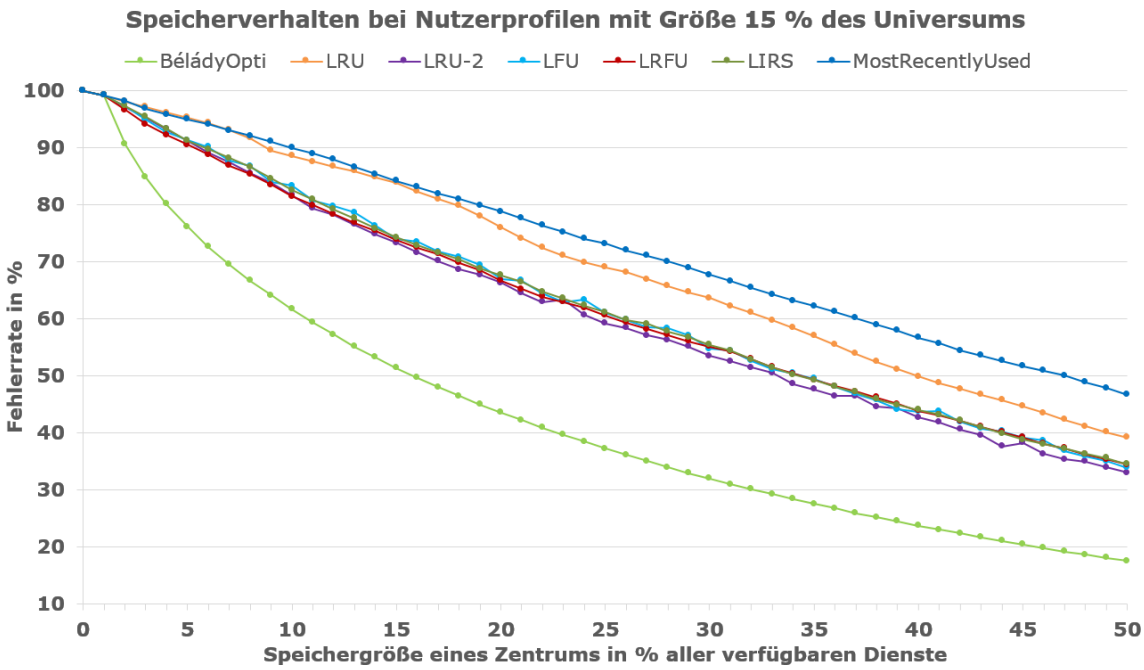


Abbildung 6.24: Leistungsvergleich von Speicherersetzungsstrategien gegenüber dem Optimum mittels Bélády bei verschiedenen Speichergrößen und 15 % Anteil der Nutzerprofile vom Universum

6.7.2 Analyse des Speicherverhaltens der Zentren und Einfluss der Nutzerprofile

In diesem Abschnitt erfolgt eine Analyse des Speicherverhaltens der Zentren in Relation zur Vielfältigkeit der Nutzerprofile. Den Nutzern wird dazu im Ansatz stets das gleiche Profil zugewiesen, wobei die Menge angeforderter Dienste verändert wird. Weiterhin erfolgt eine Variation der spezifizierten Kapazität des Speichers eines Zentrums. Alle Zentren besitzen die gleiche Kapazität. Die Art der Analyse ist auf Grund der Komplexität der Problemstellung spezifisch für ein Szenario zu erstellen. Die maßgeblichen Parameter der Simulation sind in der Tabelle 6.13 aufgeführt.

Tabelle 6.13: Parameter des Experiments Speicherverhalten

Parameter	Variable
Szenario	Index 23 - Dialtelecom (Cz)
Platzierungsstrategie	Dragoon
Entfernungsfunktion	Dijkstra mit Metrik Hops
Zielkriterium	Reduktion der maximalen Entfernung
Dienste im Universum	100
Speichergröße am Zentrum	10 % und 15 % des Universums
Anteil der Nutzerprofile	10 %, 15 % und 20 % des Universums
Zipf-Verteilung	$\alpha = 0{,}75$
Anfragen jedes Nutzers	100

Die Simulation wird gemäß dem rekursiven Systemverhalten von Abschnitt 5.1.12 durchgeführt. Ein im Zentrum gespeicherter Dienst lässt sich in dieser Simulation beliebig

oft vervielfältigen und abfragen, sodass hierfür keine Mengen zu beachten sind. Die Abbildungen 6.25 und 6.26 zeigen die Zugriffsfehlerrate bei optimaler Verwaltungsstrategie gemäß Bélády beziehungsweise bei Verwendung von der realistischen Strategie LRU-2. Hierbei sind für verschiedene Verhältnisse von Größe der Nutzerprofile und Speicherkapazität des Zentrums gegenüber gestellt.

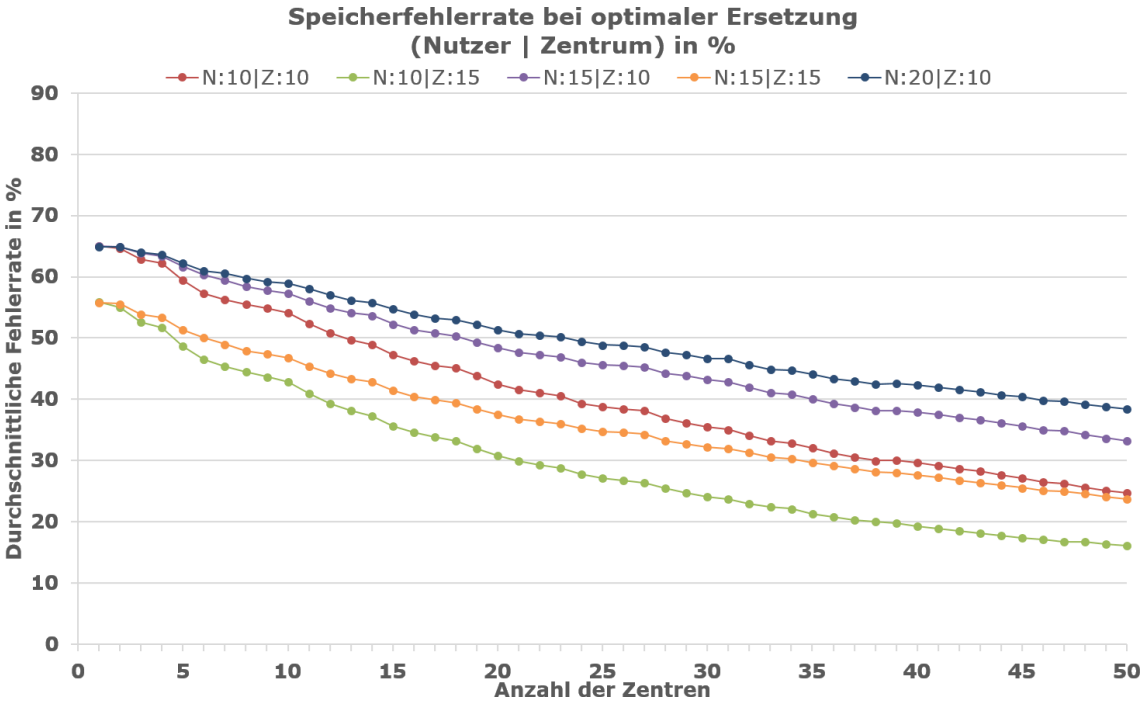


Abbildung 6.25: Einfluss der Speicherkapazität auf die Fehlerrate bei der Speicherverwaltung mittels Bélády

Auswertung:

Allgemein verringert sich mit steigender Anzahl von Zentren im System die Fehlerrate bei Anfragen. Nichtsdestotrotz verbessert nicht jedes zusätzliche Zentrum das Verhalten des Systems. Ein zusätzliches Zentrum kann bei optimierter Platzierung zu einer ungünstigeren Nutzerzuweisung führen, sodass es zu einer negativen Korrelation des Anfrageverhaltens und damit Interessen-Profilen kommt. Die Auswirkungen werden über die unterschiedlichen Anfragen am Zentrum durch die erhöhte Fehlerrate ersichtlich. Dieser Effekt ist vor allem bei geringer Anzahl von Zentren und Einsatz realistischer Speicherverwaltungsstrategien sichtbar. Zusammenfassend ergibt sich das zu erwartende Verhalten, dass sich mit steigender Größe des Speichers die Fehlerrate bei den Zentren verringert. Zudem führen kleinere Interessen-Profile zu geringeren Fehlerraten, da sich die Speicher durch die verminderte Diversität der Anfragen mittels der Verwaltungsstrategien effizienter nutzen lassen. Die Analyse ermöglicht darüber die Feststellung der notwendigen Speicherkapazität eines Zentrums in Abhängigkeit der Anzahl von Zentren, damit eine effiziente Funktionsweise des Gesamtsystems gegeben ist.

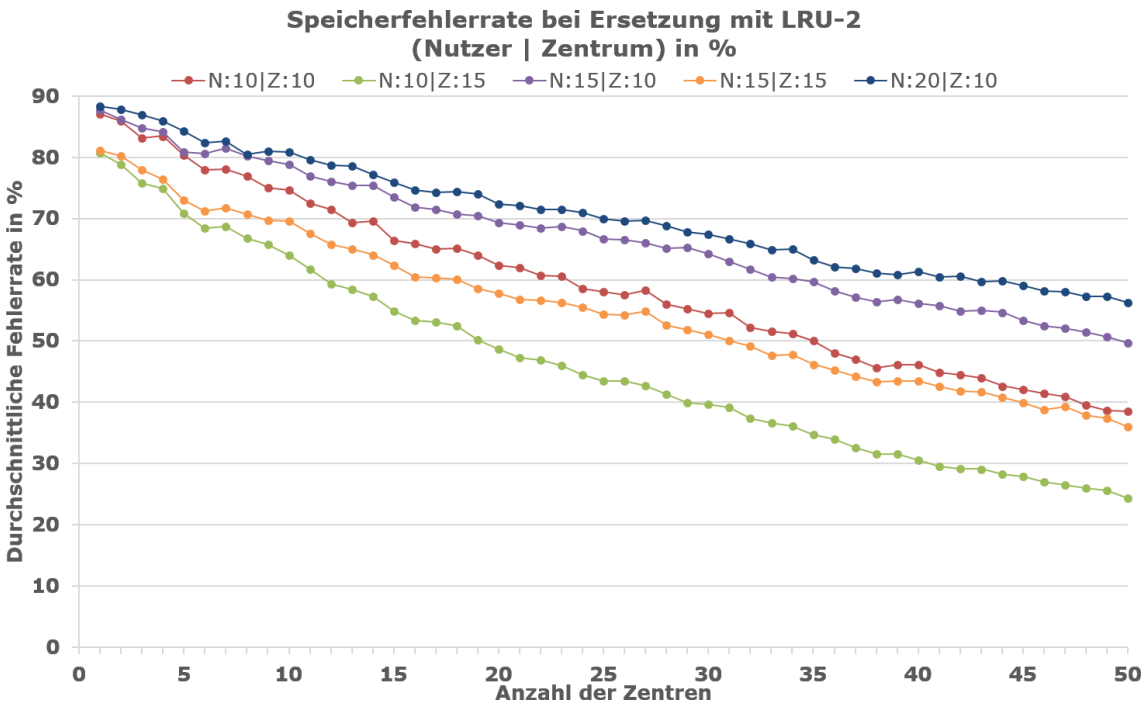


Abbildung 6.26: Einfluss der Speicherkapazität auf die Fehlerrate bei der Speicherverwaltung mittels LRU-2

6.7.3 Leistungspotential zwischen optimierter Platzierung und optimierter Zuweisung

In der folgenden Simulation wird der Leistungsunterschied zwischen optimierten Lösungen hinsichtlich verschiedener Zielkriterien aufgezeigt und welchen Einfluss dies wiederum auf das jeweilige andere Bewertungskriterium hat. Insbesondere erfolgt der Vergleich von Ergebnissen optimierter Platzierung, wobei die Entfernungen zwischen Nutzern und dem jeweiligem Zentrum reduziert sind, gegenüber Zuständen hoher Korrelation von Nutzerprofilen an den Zentren, bei denen die Entfernungen eine nachgelagerte Zielstellung darstellen. Die Simulation wird dazu mit folgenden Parametern gemäß der Tabelle 6.14 durchgeführt.

Die Abbildung 6.27 und die Tabelle 6.15 verdeutlichen den Zusammenhang von Korrelation der Nutzerprofile sowie der damit einhergehenden Anfragen an den Zentren und der Auswirkung auf die Fehlerrate bei der Zwischenspeicherung von Diensten während der Simulation. Dies erfolgt anhand verschiedener Anzahlen von Zentren im System. Der Leistungsunterschied wird insbesondere durch Lösungen ersichtlich, die nach den unterschiedlichen Optimierungskriterien Korrelation des Profils oder Platzierung ausgerichtet sind.

Auswertung:
Die Kurven und der grüne Doppelpfeil zeigen das Potential für mögliche Zwischenlösungen bei der Planung und Umsetzung des komplexen Systems auf. Dies ist vor allem für die Betreiber von der Infrastruktur und der Transportkette von Bedeutung. Die Optimierung hinsichtlich der Korrelation von den Nutzerprofilen führt zu einer geringeren Speicherfehlerrate bei Anfragen der Nutzer an die Zentren als bei einer optimierten

Tabelle 6.14: Parameter des Experiments Leistungspotential

Parameter	Variable
Szenario	Index 23 - Dialtelecom (Cz)
Platzierungsstrategie	Dragoon
Entfernungsfunktion	Dijkstra mit Metrik Hops
Zuweisungsstrategie	Greedy-Random-Walk
Korrelationsfunktion	Spearman
Zielkriterium Platzierung	Reduktion der durchschnittlichen Entfernung
Zielkriterium Korrelation	Hohe durchschnittliche Korrelation der Nutzerprofile
Dienste im Universum	100
Speichergröße am Zentrum	15 % des Universums
Anteil der Nutzerprofile	15 % des Universums
Zipf-Verteilung	$\alpha = 0{,}75$
Anfragen jedes Nutzers	100
Speicherverwaltungsstrategie	Bélády

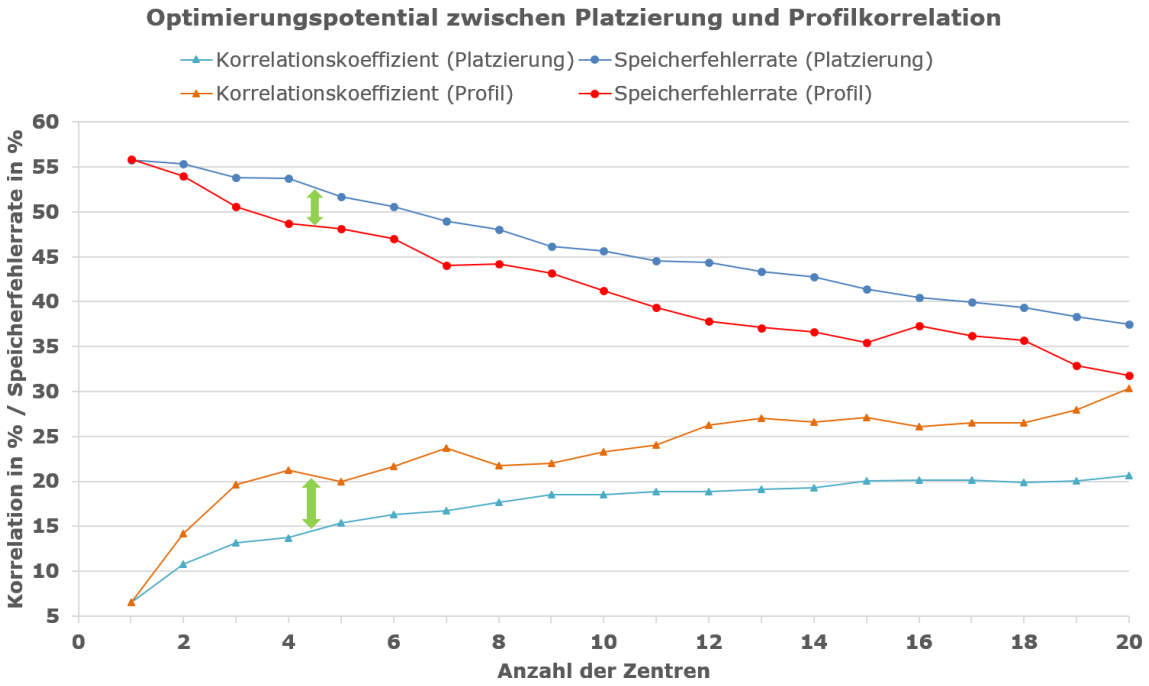


Abbildung 6.27: Zusammenhang von Korrelation der Nutzerprofile und der Fehlerrate bei Anfragen

Tabelle 6.15: Gegenüberstellung der Leistungswerte bei den Zielen kurzer Entfernung und hoher Korrelation bei 20 Zentren (Alle Angaben in %; Ein Fehler bei der Anfrage wirkt sich mit 100 % auf die Anfrageentfernung aus.)

Eigenschaft	Ziel: kurze Entfernung	Ziel: hohe Korrelation
Durchschnittliche Entfernung	3,3	22,8
Durchschnittliche Korrelation	20,6	30,3
Durchschnittliche Fehlerrate bei Anfragen	37,5	31,8
Durchschnittliche Anfrageentfernung	39,6	47,3

Platzierung. Die verschiedenen Optimierungsziele beeinflussen somit die QoS des Gesamtsystems hinsichtlich der Beantwortung der Anfragen von den Nutzern und der Ressourcenlast für den Betreiber. Unabhängig des Optimierungskriteriums kommt es bei beiden Lösungen zu einer Stagnation des Korrelationskoeffizienten mit steigender Anzahl an Zentren. Es kann dabei sogar wieder zu einer Verringerung kommen, wobei dies auf die Komplexität des Problems und der Leistung von den heuristischen Optimierungsalgorithmen zurückzuführen ist. Nichtsdestotrotz bleibt der Einfluss auf die Fehlerrate in dieser Situation gering, da sich die erhöhte Speicherkapazität mit jedem zusätzlichen Zentrum im gesamten System positiver auswirkt und ein verbessertes Speicherverhalten eintritt. Weiterhin führt eine verbesserte Profilkorrelation nicht automatisch zu einer verbesserten Speicherfehlerrate während der Simulation. Dies liegt zum einen an dem stochastischen Anfrageverhalten der Nutzer als auch an der Speichergröße der Zentren. Die Optimierung hinsichtlich einer besseren Profilkorrelation anstatt der vorteilhaften Platzierung führt zu einer höheren durchschnittlichen Entfernung von Nutzer und Zentrum, welche zudem stärker variiert. Diese Informationen unterstützen einen Dienstanbieter bei Entscheidungen über zusätzliche Zentren zur Verbesserung des Gesamtsystems und der erreichbaren QoS.

6.7.4 Pareto-Front

Bei der Entscheidungsunterstützung für ein spezifisches Szenario sind mehrere Lösungsmöglichkeiten zur Auswahl gefragt, insbesondere bei multikriteriellen Problemstellungen. Dazu wird in der folgenden Optimierung gezeigt, dass sich über den generischen Ansatz diese Lösungen berechnen lassen. Hierbei wird die Optimierung hinsichtlich der beiden Kriterien geringer durchschnittlicher Entfernung zwischen Zentrum und Nutzer sowie einer hohen Korrelation von Interessen-Profilen an den Zentren vorgenommen. Grundlage bildet das Testsetup gemäß Tabelle 6.16. Die jeweils besten Lösungen der Algorithmen bilden die Pareto-Front, welche in den Abbildungen 6.28 und 6.29 für das Szenario mit 5 und 10 Zentren dargestellt sind. Die Algorithmen NSGAII, MOCell, PAES, SPEA2 und VEGA sind von dem Framework MOEA, wobei diese über einen Adapter mit einer Liste von Parametern der Permutation angebunden werden. Der erste Parameter bildet dabei die knotenbasierte Platzierung der Zentren ab und der zweite Parameter gibt die Zuweisung der Nutzer zu den Zentren an. Weiterhin wird eine Re-Implementierung des AMOSA Verfahrens eingesetzt, wobei dieses bei der Initialisierung Lösungen von Zwei-Approx, Greedy und Dragoon verwendet. Zudem wird zum Vergleich der sequentielle Optimierungsansatz aus Kapitel 5.1.11 eingesetzt, wobei die initiale Lösung mit dem Ziel der minimierten Entfernung durch Dragoon entsteht. Anschließend übernimmt ein Random Walk Verfahren die Optimierung hinsichtlich einer hohen Korrelation.

Auswertung:
Der grüne Pfeil veranschaulicht das Optimierungsziel. Die unterschiedlichen Ausprägungen und Abstände zwischen den Pareto-Fronten verdeutlichen die Komplexität der Problemstellung. Durch die Referenz mit der sequentiellen Optimierung mittels Dragoon und Random Walk wird deutlich, dass die Optimierungsverfahren des Frameworks MOEA mit der verwendeten Repräsentation für diese Problemstellung ungeeignet sind. Der Grund dafür ist in der unflexiblen Anwendbarkeit des Frameworks MOEA zu sehen. Mit

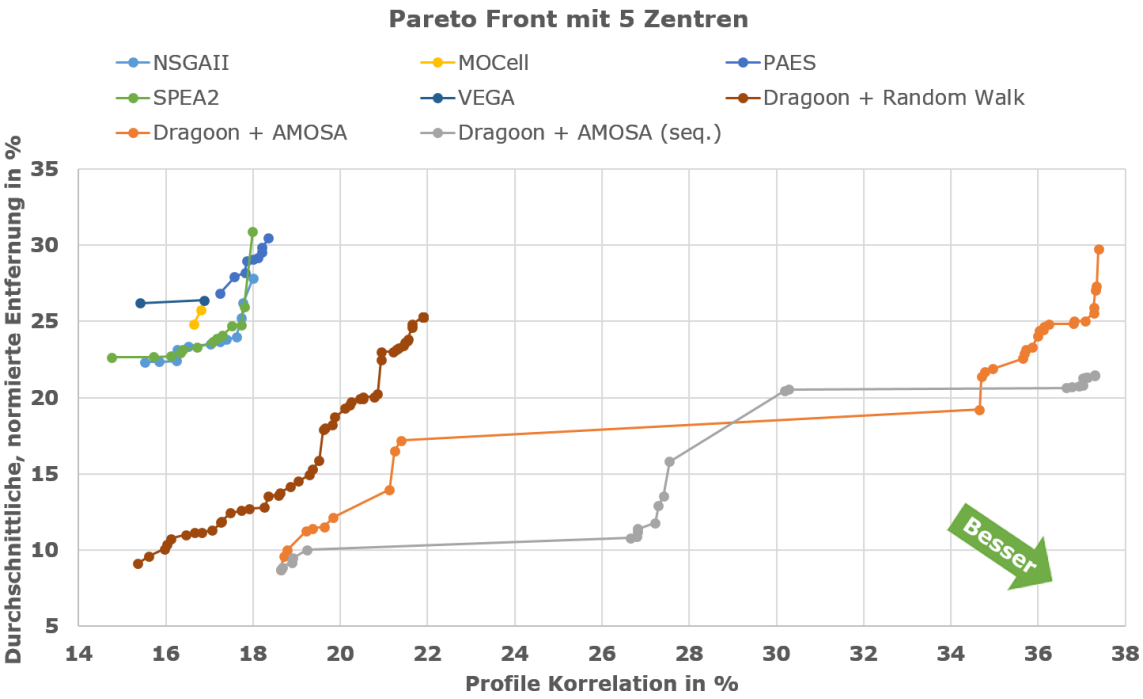


Abbildung 6.28: Pareto-Front bei der Optimierung von der Entfernung und der Profilkorrelation mit 5 Zentren

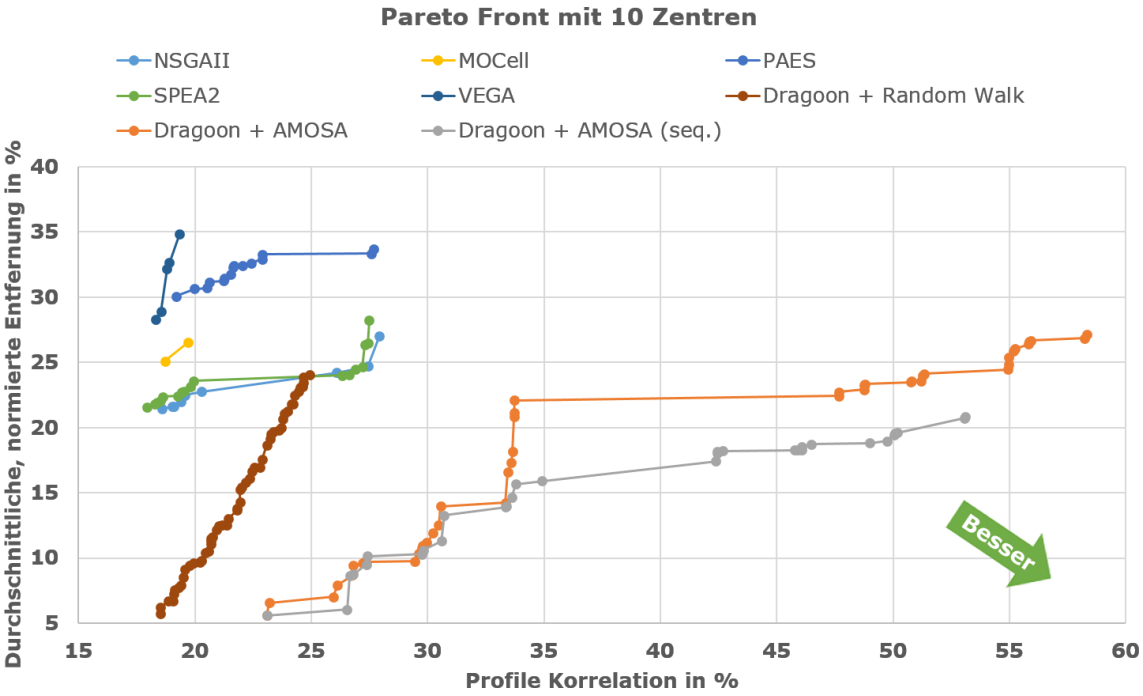


Abbildung 6.29: Pareto-Front bei der Optimierung von der Entfernung und der Profilkorrelation mit 10 Zentren

Tabelle 6.16: Parameter des Experiments Pareto-Front

Parameter	Variable
Szenario	Index 23 - Dialtelecom (Cz)
Entfernungsfunktion	Dijkstra mit Metrik Hops
Korrelationsfunktion	Spearman
Zielkriterium Platzierung	Reduktion der durchschnittlichen Entfernung
Zielkriterium Korrelation	Hohe durchschnittliche Korrelation der Nutzerprofile
Dienste im Universum	100
Speichergröße am Zentrum	15 % des Universums
Anteil der Nutzerprofile	15 % des Universums
Zipf-Verteilung	$\alpha = 0{,}75$

einer spezifischeren Anpassung sind auch bei diesen Verfahren bessere Ergebnisse zu erwarten. Der angepasste Algorithmus AMOSA erreicht mit der initialisierten Lösung von Dragoon in dem Szenario die beste Leistung. Insbesondere ist die Leistung des sequentiell vorgehenden AMOSA Ansatzes zu beachten. Hierbei führt der Algorithmus eine abwechselnde Optimierung in jeder Iteration aus, sodass entweder nur die Platzierung verbessert wird oder eine Optimierung der Zuweisung erfolgt. Insgesamt zeigt diese Analyse, dass sich auf Grundlage des generischen Modells Zwischenlösungen hinsichtlich mehrerer Zielkriterien über eine multikriterielle Optimierung berechnen lassen. Durch die Variation der Parameter werden einem Entscheidungsträger die Leistungsunterschiede der Lösungen verdeutlicht, insbesondere im Zusammenhang mit der Simulation lassen sich somit die Auswirkungen auf die QoS und QoE schlussfolgern.

7 Zusammenfassung und Ausblick

Das Ziel der vorliegenden Dissertation ist die Entwicklung eines generischen Modells für Transportketten zur Planung und Optimierung von Verteilzentren. Beginnend mit einer Einführung in das Thema und der Einordnung der Arbeit folgte in Kapitel 1.2 die Erläuterung der Anwendungsgebiete mit Beispielen realer komplexer Systeme. Der Fokus liegt dabei auf zweistufigen Transportketten als kürzeste Kette mit Verteilzentren zur Zwischenspeicherung, wobei diese ein grundlegendes Element des Transports sind. Dazu wurden Szenarien aus den Bereichen der Logistik mit dem *Facility Location Problem* und Rechnernetze mit *Content Delivery Network* vorgestellt. Durch eine vergleichende Analyse wird die Motivation für einen einheitlichen Ansatz ersichtlich. In Anlehnung daran werden in Kapitel 1.3 wissenschaftliche Fragestellungen aufgestellt und einzeln näher erläutert. Die Zielstellung ist die Entwicklung eines generischen Modells für Transportketten gleichen Funktionsprinzips mit flexibler Anpassbarkeit an die Spezifika der betrachteten Szenarien. Zentraler Ausgangspunkt ist das k-center und k-cluster Problem. In Kapitel 2 wird dazu die Problemstellung im Detail beschrieben und die erforderlichen Bestandteile einer adäquaten Lösung aufgeführt. Diese besteht aus den drei Bestandteilen einer generischen Modellierung, einer konfigurierbaren Simulation und einer optimierten Lösungsfindung. Weiterhin wurden hier die verschiedenen Herausforderungen und erforderlichen Lösungsprozesse einzeln analysiert und deren Zusammenhänge dargestellt. Insbesondere sind während der Optimierung die sich wechselseitig beeinflussenden Teilprobleme für eine ganzheitliche Lösung gleichzeitig zu betrachten.

Darauf aufbauend wurden in Kapitel 3 die Anforderungen für ein generisches Modell abgeleitet, welche sich in funktional und nicht-funktional unterteilen. Darüber hinaus erfolgt eine Analyse bezüglich der Struktur für einen Lösungsansatz, um die Herausforderungen systemübergreifender Aspekte zu beachten. Zur Bewertung des Stands der Wissenschaft wurde entsprechend der Anforderungen eine Liste mit Bewertungskriterien erstellt. Anhand dieser Kriterien sind in Kapitel 4 an die Aufgabenstellung angrenzende Arbeiten beschrieben. Das Kapitel unterteilt sich in mehrere Abschnitte entsprechend der betreffenden Fachgebiete Logistik, Rechnernetze und ganzheitlicher Ansätze mit domänenübergreifender Betrachtungsweise. Zu bestehenden Ansätzen wurde individuell Bezug genommen und die Limitierungen verdeutlicht. Anschließend erfolgt in Abschnitt 4.6 eine Zusammenfassung bisheriger Arbeiten. Hierbei wird das Fehlen eines umfassenden wissenschaftlichen Ansatzes für das Problem ersichtlich und der Mehrwert zur vorliegenden Arbeit aufgezeigt.

In Kapitel 5 wird das Konzept eines generischen Modells von zweistufigen Transportketten vorgestellt und es werden die Prozesse zur Lösungsberechnung erklärt. Im Anschluss erfolgt eine Beschreibung zur Adaption auf die Anwendungsbereiche der Logistik,

Rechnernetze und Geolokalisation. Zur Berechnung optimierter Lösungen folgt die Erläuterung des entwickelten Algorithmus Dragoon für die unterschiedlichen Bedingungen der Platzierung von Zentren.

Daran anschließend beschreibt Kapitel 6 die Implementierung des generischen Modells. Dadurch wird das Ergebnis des Konzeptes verifiziert. Ein essentieller Bestandteil ist die Lösungsfindung zur Platzierung der Zentren. Das Ergebnis der Optimierung beeinflusst maßgeblich die Leistung des Gesamtsystems. Daher erfolgt zunächst eine Analyse zur Leistungsfähigkeit des entwickelten Algorithmus Dragoon gegenüber bestehenden Ansätzen im Bereich der Platzierung, wobei dessen Potential deutlich wird. Der Algorithmus liefert im Durchschnitt bessere Ergebnisse als die Referenzansätze nahe dem globalen Optimum. Dabei wurde der Einfluss von Zufall gegenüber bisherigen Clusterverfahren reduziert und somit der Berechnungsaufwand gemindert. Die Simulationen zeigen die Tragfähigkeit des Modells und lassen eine Analyse hinsichtlich der verschiedenen Einflussparameter zu. Insbesondere werden die Fragen zur Anzahl an Zentren, deren Platzierung und der Zuweisung von Nutzern sowie der Speichergröße beantwortet. Zudem wird durch die Berechnung der Pareto-Front mittels multikriterieller Optimierung eine Auswahl von optimierten Lösungen gegeben.

7.1 Beantwortung der wissenschaftlichen Fragestellungen

In Bezug zu den wissenschaftlichen Fragestellungen aus Kapitel 1.3 ergeben sich folgende zusammengefasste Antworten.

Fragestellung 1: Was ist ein multikriterieller Optimierungsansatz, der für verschiedene Domänen einer Transportkette anwendbar ist?

Diese Fragestellung betrachtet die Modellierung, Simulation und Optimierung von zweistufigen Transportketten unterschiedlicher Domänen mit einem einheitlichen Ansatz. Durch die Abstraktion verschiedener Anwendungsszenarien auf das k-center und das k-cluster Problem wird die Grundlage für einen gemeinsamen Lösungsansatz geschaffen. Diese kombinatorischen Optimierungsprobleme bilden die Basis zur Modellierung des generischen Modells. Die Umsetzung erfolgte entsprechend in Bezug auf zweistufige Transportketten. Mit der Parametrierung des Modells sowie der spezifischen Anpassbarkeit von Nebenbedingungen, Zielkriterien und Optimierungsansätze ergibt sich ein einheitlicher Ansatz für die unterschiedlichen Eingangsdaten der Szenarien. Weiterhin lässt sich auf Grund dessen das Modell und die verschiedenen Nebenbedingungen in weiteren Prozessschritten verwenden, sodass diese in der Optimierung und Simulation Anwendung finden können.

Fragestellung 2: Wie lassen sich Verhaltensprofile von Nutzern mit in die Planung und Optimierung einbeziehen, um die Leistung des Gesamtsystems zu verbessern?

Bei der zweiten wissenschaftlichen Fragestellung ist zu klären, wie sich Kenntnisse über Nutzerverhalten vorteilhaft in die Planung des Systems einbeziehen lassen. Dazu erfolgt während der Optimierung und Simulation eine zielgerichtete Zuweisung der

Nutzeranfragen zu spezifischen Zentren. Bei diesen ist der angefragte Dienst mit erhöhter Wahrscheinlichkeit sofort verfügbar. Dafür werden die Korrelationskoeffizienten der Profile bei der Zuweisung von Nutzern berücksichtigt und hinsichtlich einer positiven Korrelation optimiert.

Fragestellung 3: Wie lassen sich verschiedene Optimierungsaspekte der Zentren, welche das Gesamtsystem beeinflussen, für eine integrierte Lösung parallel betrachten?

Die dritte wissenschaftliche Fragestellung zielt auf eine ganzheitliche, teilproblemübergreifende Optimierung des Systems. Unter Nutzung bestehender sowie neu entwickelter Optimierungsansätze auf die durch das generische Modell repräsentierte Problemstellung ergeben sich verschiedene Lösungsmöglichkeiten.

Dazu wurden die Algorithmen entsprechend flexibilisiert und adaptiert, um deren Anwendbarkeit bei unterschiedlichen Nebenbedingungen sicher zu stellen. Weiterhin lassen sich durch das Modell verschiedene Zielkriterien den Algorithmen vorgeben, sodass eine übergreifende Optimierung ermöglicht wird. Über die Berechnung der Pareto-Front mit mehreren Kriterien ergeben sich für kooperierende Systembetreiber hinsichtlich einer Zusammenarbeit integrierte und angepasste Lösungen. Im Zusammenhang mit der Simulation grundlegender Prozesse lassen sich darüber hinausgehend operative Auswirkungen beurteilen und verschiedene Parameter und Strategien analysieren.

7.2 Bewertung des generischen Modells und des Optimierungsansatzes

Das generische Modell deckt mehrere Anwendungsbereiche ab und bietet sich damit als Grundlage für spezifische Weiterentwicklungen an. Dabei erfolgt zudem eine Modellierung von Interessen-Profilen der Nutzer, wodurch eine Analyse von stochastischen Einflüssen in die weiteren Berechnungen ermöglicht wird. Der beschriebene Ansatz erlaubt zudem eine Berücksichtigung der teilproblemübergreifenden Zusammenhänge, wodurch eine bessere Lösung für das Gesamtsystem erreicht wird als bei separater Betrachtung. Zusätzlich erfolgt die Einbeziehung angrenzender Systeme und die Beachtung ihrer Auswirkungen, sodass eine angepasste Integration realisierbar ist. Die Modellierung der verschiedenen Optimierungsaspekte lässt dabei eine problemübergreifende Analyse hinsichtlich mehrerer Zielkriterien zu. Zur Lösungsfindung wurde ein neuer Algorithmus zur Platzierung entwickelt, welcher im Durchschnitt für die betrachteten Szenarien bessere Ergebnisse erreicht als bestehende Verfahren. Weiterhin dient das generische Modell für eine multikriterielle Optimierung, wobei sich mehrere Zielkriterien gleichzeitig in Betracht ziehen lassen. Dies ermöglicht die Berechnung von Zwischenlösungen. Über die Ermittlung der Pareto-Front lassen sich optimierte Lösungen für eine kooperative Zusammenarbeit mehrerer Betreiber wie von Infrastruktur, Verteilung und Dienstanbieter auswählen. Bei entsprechend gegenüber einander ausbalancierten Lösungen führen diese zu einer gesteigerten Dienstqualität, sodass letztendlich der Nutzer davon profitiert. Der beschriebene Ansatz gewährleistet die Simulation auf Basis des generischen Modells unter Verwendung flexibler Szenariendaten. Dabei lassen sich die parametrisierbaren

Nebenbedingungen mit integrieren. In den komplexen Planungsprozess werden die Nutzerprofile mit einbezogen, wodurch sich eine ganzheitliche Lösung berechnen lässt. Dies führt zu einer bedarfsorientierten Systemplanung und ermöglicht Analysen über die Auswirkungen bei Veränderungen am System.

7.3 Zukünftige Forschungsbereiche

Das generische Modell bietet mehrere Möglichkeiten zur Weiterentwicklung hinsichtlich anderer Anwendungsbereiche und zur Verbesserung bestehender Lösungsmethoden. Darüber hinaus besteht das Potential zur spezifischen Anpassung an einzelne Szenarien im Bereich der Logistik oder Rechnernetze, wobei die Abbildung mit einem erhöhten Detailgrad erfolgt. Zudem ist eine zukünftige Entwicklung hin zu wirtschaftswissenschaftlichen Ökosystemen bei Supply Chains möglich. Dabei wird jeder Nutzer gleichzeitig auch zu einem Anbieter von Diensten, wodurch ein Wettbewerb innerhalb kooperierender Partner entsteht. Hierbei verändert sich die Transportkette von einer linearen Struktur zu einem komplexen, dynamischen Netz mit proaktiver Versorgung. Basierend auf dem generischen Modell ist für ein solches unstrukturiertes System die Simulation und Optimierung entsprechend zu erweitern.

Im Hinblick auf zukünftige Entwicklungen und neuer Optimierungsverfahren sind diese zur Leistungsverbesserung insbesondere im Bereich multikriterieller Optimierung zu evaluieren. Weiterhin ist die Funktionalität der Datenverarbeitung von Szenarien und die Präsentation der Lösungen benutzerfreundlicher zu gestalten. In Abhängigkeit des Szenarios gestattet das vorgestellte Modell eine Optimierung der Infrastruktur vorzunehmen. Dabei werden die Verbindungen der einzelnen Teilnehmer genauer analysiert und gegebenenfalls neue Verbindungen vorgeschlagen, um die Last zu verteilen und die Dienstqualität zu verbessern. Zu diesen allgemeinen Möglichkeiten der Weiterentwicklung werden im Folgenden gezielte, wissenschaftliche Vertiefungen aufgezeigt.

Einbeziehen von Lastausgleich

Als weitere Forschungsrichtung bietet sich die Konzeption und Implementierung von Lastausgleichsstrategien an. Zum einen lassen sich diese Verfahren analysieren und die Auswirkungen im Zusammenhang mit diversen Einflüssen betrachten. Zum anderen ermöglichen Lastausgleichsstrategien die Funktionsweise bei Lastspitzen auf die Infrastruktur und die Zentren mittels Simulation genauer zu bewerten. In weiteren wissenschaftlichen Arbeiten sind dazu spezifische Mechanismen zur Behandlung von Ausfällen zu entwickeln. Dies Thematik ist im Zusammenhang mit einer redundant ausgelegten Platzierung der Zentren zu analysieren.

Weitere Dynamisierung des Systems

Eine zukünftige Entwicklungsrichtung ist die weitere Dynamisierung des Systems. Hierbei sind zur Erschließung weiterer Anwendungsgebiete sich verändernde Eigenschaften der Infrastruktur zu realisieren. Dazu gehören unter anderem die Berücksichtigung zeitabhängige Einflüsse wie die gegenwärtige Auslastung durch Verkehr. Zusätzlich sind Strategien zur dynamischen Verschiebung der Zentren während des operativen Betriebs vorzusehen, um Szenarien mit langsam verändernden Anforderungsprofilen angemessen betrachten zu können. Eine Weiterentwicklung der zweistufigen Transportkette hin zu einer beliebig stufigen Kette ermöglicht eine direkte Anwendbarkeit auf längere Transportketten. Dazu ist das allgemeine Zentrum um die weitere Eigenschaft Stufe zu ergänzen. In der Simulation erfolgt daraufhin der Transfer von Diensten vorrangig zwischen Zentren aneinander grenzender Stufen. Ein Transfer zwischen Zentren gleicher Stufe verbessert zudem den Lastausgleich. Weiterhin ist das Hinzukommen oder Wegfallen von Nutzern im Modell vorzusehen. Dazu erhalten die Nutzer eine weitere Eigenschaft Aktivität, welche den aktiven Zeitraum spezifiziert. Somit verändert sich über die betrachtete Laufzeit die Menge der beeinflussenden Nutzer. In diesem Zusammenhang ist eine Veränderung der durch das System angebotenen Arten und Menge von Diensten einzuplanen, wobei diese verschiedene Eigenschaften aufweisen. Dies erfordert neue Vorgehensweisen bei der Verwaltung des Speichers an den Zentren. Die genannten dynamischen Aspekte wirken sich insbesondere auf die Simulation des operativen Betriebs aus, wodurch sich stochastische Einflüsse und Konkurrenzsituationen analysieren lassen.

Betrachtung von lokalen Ressourcen zur Verteilung der Dienste

Durch Ergänzung um verfügbare Ressourcen an einem Zentrum zur Auslieferung von Diensten lassen sich Bedarfsforderungen für einen effizienten Betrieb aufstellen. Die Spezifizierung einer Obergrenze für die maximale Anzahl bedienbarer Nutzer an einem Zentrum ergibt während der Optimierung eine geänderte Aufgabenstellung. Hierbei sind entsprechende Kapazitäten zu berücksichtigen. Dies führt sowohl in der Logistik als auch bei Rechnernetzen zu realitätsnäheren Ergebnissen. Die Einbeziehung bereits vorhandener Ressourcen wie bestehenden Zentren ermöglicht die Anwendbarkeit auf operativen Systemen zur Leistungssteigerung. Im Bereich der Logistik ergeben sich vor allem durch die zusätzliche Spezifikation von ein oder mehreren Auslieferern an einem Zentrum weitere Optimierungsmöglichkeiten. Über die Berechnung der tourbasierten Auslieferung sind optimierte Routen zur Verteilung zu ermitteln. Dabei ist zu bestimmen, welche Anzahl an Auslieferer zur effizienten Verteilung vorteilhaft ist. Dementsprechend sind die Ladekapazitäten und Personal vorauszuplanen

Vertiefung des operativen Verhaltens mit Unsicherheiten

Ein weiteres Forschungsgebiet betrifft die Zuweisung von Anfragen der Nutzer an die Zentren während des operativen Betriebs. Durch die Ergänzung um probabilistisches Verhalten wird keine eindeutige Zuordnung vorgegeben. Dabei bestimmt eine Wahr-

scheinlichkeitsauswahl das Zentrum für die Anfrage eines Nutzers. Somit lassen sich Ungenauigkeiten in den Ausgangsdaten mit in die Simulation einbeziehen und Varianzen im Nachfrageverhalten abbilden. Zudem sind verschiedene Zuweisungsstrategien zu evaluieren, z. B. an welches Zentrum der Nutzer seine initiale Anfrage zu stellen hat und von welchem Zentrum die Anfrage anschließend bearbeitet wird. Dabei ist abzuwägen, ob weiter entfernte Zentren mit sofortiger Verfügbarkeit die Bearbeitung übernehmen oder näher gelegene Zentren eine vorherige Zwischenspeicherung vornehmen. Eine Speicherung bietet weiteren Nutzern mit demselben Interesse eine sofortige Verfügbarkeit, wobei wiederum dafür andere Dienste aus dem Speicher verdrängt werden. Dies betrifft auch weitere Strategien zur Unterteilung des Speichers in persistente und dynamische Bereiche. Die im persistenten Anteil des Speichers vorgehaltenen Dienste werden dauerhaft vom Zentrum angeboten, wohingegen der dynamische Speicher entsprechend der Anfragen mit den angeforderten Diensten belegt wird.

Entsprechend der aufgezeigten, zukünftigen Forschungsbereiche stellt das Ergebnis dieser Arbeit zusätzlich zu den erreichten Resultaten eine Plattform dar. Diese lässt sich in mehreren Anwendungsgebiete einsetzen und ist eine Basis für vielfältige Weiterentwicklungen.

Literaturverzeichnis

[1] B. Bojesen, “Dänische Szene,” *Newspaper: Politiken*, 1967.

[2] J. Krarup and P. Pruzan, “Selected families of location problems,” in *Discrete Optimization and Systems Applications of the Systems Science Panel of NATO and of the Discrete Optimization Symposium* (E. J. P.L. Hammer and B. Korte, eds.), vol. 5 of *Annals of Discrete Mathematics*, pp. 327 – 333, Elsevier, 1979.

[3] T. Drezner and H. A. Eiselt, *Facility Location: Applications and Theory*, ch. 5 Consumers in Competitive Location Models. Springer-Verlag Berlin Heidelberg, 2002.

[4] T. Drezner and Z. Drezner, “The gravity p-median model,” *European Journal of Operational Research*, vol. 179, no. 3, pp. 1239 – 1251, 2007.

[5] K. Haase, S. Müller, and M. Hoppe, “Kunden beeinflussen die betriebliche Standortplanung,” *Logistik - Wissenschaftliche Zeitschrift der Technischen Universität Dresden*, vol. 58, pp. 27–32, 2010.

[6] A. Schrijver, “On the history of combinatorial optimization (till 1960),” *Handbooks in Operations Research and Management Science*, vol. 12, pp. 1–68, 2005.

[7] Ernst & Young, “Der Fels in der Brandung? - Standort Deutschland,” 2012.

[8] Statistisches Bundesamt, “Verkehr - Fachserie 8, Reihe 1.1,” 2017.

[9] Cisco Systems, “Cisco Visual Networking Index,” 2016.

[10] G. Stiller, “Transportkette.” `http://www.wirtschaftslexikon24.com/d/transportkette/transportkette.htm` (2016-05-20), 2015.

[11] DIN 30781-1:1989-05, *Transportkette; Grundbegriffe*, 1989.

[12] DIN 30781-2:1989-05, *Transportkette; Systematik der Transportmittel und Transportwege*, 1989.

[13] T. Gudehus, *Logistik 2: Netzwerke, Systeme und Lieferketten*. Springer VDI, 2000.

[14] G. Stiller, “Standortwahl.” `http://www.wirtschaftslexikon24.com/d/standortwahl/standortwahl.htm` (2016-05-24), 2015.

[15] F. Lehrieder, *Performance Evaluation and Optimization of Content Distribution using Overlay Networks*. PhD thesis, Universität Würzburg, Fakultät für Mathematik und Informatik, 2013.

[16] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica, "A Data-oriented (and Beyond) Network Architecture," *SIGCOMM Computer Communication Review*, vol. 37, pp. 181–192, Aug. 2007.

[17] R. Gomes, J. P. de Sousa, and T. G. ao Dias, "A grasp-based approach for demand responsive transportation," in *International Journal of Transportation*, vol. 2, pp. 21–32, 2014.

[18] Y. Xiang, F. Hou-ming, G. Zheng-feng, and L. Yang, "Modelling of mixed hub-and-spoke express freight network considering customer satisfaction," in *IEEE International Conference on Computer and Communications (ICCC)*, pp. 2481–2485, 2016.

[19] S. Knight, H. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The Internet Topology Zoo," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1765–1775, 2011. `http://topology-zoo.org/` (2017-02-14).

[20] V. Aggarwal, A. Feldmann, and C. Scheideler, "Can ISPs and P2P Users Cooperate for Improved Performance?," in *Computer Communication Review*, vol. 37 of *3*, pp. 29–40, ACM SIGCOMM, 2007.

[21] B. Briscoe, A. Brunstrom, A. Petlund, D. Hayes, D. Ros, I. J. Tsang, S. Gjessing, G. Fairhurst, C. Griwodz, and M. Welzl, "Reducing internet latency: A survey of techniques and their merits," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2149–2196, 2016.

[22] V. Aggarwal, O. Akonjang, and A. Feldmann, "Improving user and ISP experience through ISP-aided P2P locality," in *IEEE International Conference on Computer Communications (INFOCOM) Workshops*, pp. 1–6, 2008.

[23] L. C. Shu, L. Shu, and M. Young, "An Overview of Real-Time Concurrency Control with Analytic Worst-Case Latency Guarantees," in *IEEE Workshop on Real-Time Operating Systems and Software*, pp. 69–73, 1993.

[24] I. Melve, "Why internet service providers should integrate web caches into their networks," *Web Caching in the Internet Conference*, 1996.

[25] C. Dvorak, *Definition of Quality of Experience (QoE)*. Telecommunication Standardization Sector of ITU, 2007. `https://www.itu.int/md/T05-FG.IPTV-IL-0050/en` (2017-02-14).

[26] M. J. Freedman, E. Freudenthal, and D. Mazières, "Democratizing Content Publication with Coral," in *Symposium on Networked Systems Design and Implementation*, vol. 1 of *NSDI*, pp. 1–14, USENIX Association, 2004.

[27] P. Pöml, "MirrorBrain." `http://mirrorbrain.org/` (2017-02-14), 2008.

[28] G. Kimchi, D. Buchmueller, S. A. Green, B. C. Beckman, S. Isaacs, A. Navot, F. Hensel, A. Bar-Zeev, S. Sylvain, and J.-M. Rault, *Unmanned Aerial Vehicle Delivery System*, 2015. Amazon Technologies, Inc., Patent US 2015/0120094 A1.

[29] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. Duarte, “Improving the accuracy of measurement-based geographic location of Internet hosts,” *Computer Networks*, vol. 47, no. 4, pp. 503 – 523, 2005.

[30] X. Qin, P. Xia, C. Ngo, H. Singh, H. R. Shao, H. Niu, C. Kwon, G. Fan, and S. Kim, “Autonomous Coordinator Selection in Beamformed 60GHz Wireless Networks,” in *IEEE Wireless Communications and Networking Conference*, pp. 1815–1820, 2008.

[31] T. W. Sung, L. Kong, P. W. Tsai, and J. S. Pan, “A distance coefficient-based algorithm for k-center selection in wireless sensor networks,” in *IEEE International Conference on Consumer Electronics (ICCE)*, pp. 293–294, 2017.

[32] J. Wu, D. Zhao, L. Lu, J. Tian, and J. Xiang, “A Comparative Study of Knowledge Management on Undergraduate by Questionnaire,” in *IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pp. 486–492, 2017.

[33] D. Zhou, L. Zhang, and X. Deng, “Multi-layer multi-center atom set cohesion clustering algorithm,” in *International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1–6, 2014.

[34] K. S. Sahoo, B. Sahoo, R. Dash, and M. Tiwary, “Solving Multi-controller Placement Problem in Software Defined Network,” in *IEEE International Conference on Information Technology (ICIT)*, pp. 188–192, 2016.

[35] B. Frank, I. Poese, G. Smaragdakis, A. Feldmann, B. M. Maggs, S. Uhlig, V. Aggarwal, and F. Schneider, “Collaboration Opportunities for Content Delivery and Network Infrastructures,” *ACM SIGCOMM ebook on Recent Advances in Networking*, vol. 1, 2013.

[36] M. Rodrigues, S. Fernandes, J. Kelner, and D. Sadok, “An Analytical View of Multiple CDNs Collaboration,” in *Proceedings of the International Conference on Advanced Information Networking and Applications (AINA)*, pp. 25–32, IEEE Computer Society, 2014.

[37] G. Cabrera, S. Gonzalez-Martin, A. A. Juan, J. M. Marquès, and S. E. Grasman, “Combining Biased Random Sampling with Metaheuristics for the Facility Location Problem in Distributed Computer Systems,” in *Winter Simulation Conference (WSC)*, pp. 3000–3011, 2014.

[38] D. Papadimitriou, “New challenges in network optimization,” in *IEEE International Conference on High Performance Switching and Routing (HPSR)*, pp. 1–7, 2016.

[39] M. Garetto, E. Leonardi, and V. Martina, “A Unified Approach to the Performance Analysis of Caching Systems,” *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, vol. 1, no. 3, pp. 12:1–12:28, 2016.

[40] T. Yoshizumi and H. Okano, “A simulation-based algorithm for supply chain optimization,” in *Winter Simulation Conference (WSC)*, pp. 1924–1931, 2007.

[41] J. Sahoo, M. A. Salahuddin, R. Glitho, H. Elbiaze, and W. Ajib, “A Survey on Replica Server Placement Algorithms for Content Delivery Networks,” *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 1002–1026, 2017.

[42] Internationale Norm, “Qualitätsmanagementsysteme - Grundlagen und Begriffe (DIN EN ISO 9000:2015-11),” 2015.

[43] M. Agrebi, M. Abed, and M. N. Omri, “Urban distribution centers’ location selection’s problem: A survey,” in *International Conference on Advanced Logistics and Transport (ICALT)*, pp. 246–251, 2015.

[44] A. Larsen, *The Dynamic Vehicle Routing Problem*. PhD thesis, Department of Mathematical Modelling, The Technical University of Denmark, Building 321, DTU, DK-2800 Kgs. Lyngby, 2000.

[45] H. N. Psaraftis, M. Wen, and C. A. Kontovas, “Dynamic Vehicle Routing Problems: Three Decades and Counting,” *Networks*, vol. 67, no. 1, pp. 3–31, 2016.

[46] Telecommunication Standardization Sector of ITU, *E.800 : Definitions of terms related to quality of service*, 2008. `http://www.itu.int/rec/T-REC-E.800-200809-I` (2017-02-14).

[47] Telecommunication Standardization Sector of ITU, *E.802 : Framework and methodologies for the determination and application of QoS parameters*, 2007. `https://www.itu.int/rec/T-REC-E.802-200702-I/en` (2017-02-14).

[48] M. J. Geiger, *Multikriterielle Optimierung*, pp. 7–18. Wiesbaden: Deutscher Universitätsverlag, 2005.

[49] F. X. Bea, E. Dichtl, and M. Schweitzer, *Allgemeine Betriebswirtschaftslehre, Bd. 3: Leistungsprozess.*, vol. 9. Stuttgart: Lucius & Lucius - UTB, 2006.

[50] R. L. Francis, L. F. McGinnis, and J. A. White, “Locational analysis,” *European Journal of Operational Research*, vol. 12, no. 3, pp. 220–252, 1983.

[51] O. Haaker, *Standortwahl von internationalen Industrieunternehmen: Eine modell-theoretische Analyse*. Springer Gabler, 2014.

[52] S. J. Jafari and H. Naji, “GeoIP Clustering: Solving Replica Server Placement problem in Content Delivery Networks by clustering users according to their physical locations,” in *Conference on Information and Knowledge Technology (IKT)*, pp. 502–507, 2013.

[53] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.

[54] T. F. Gonzalez, “Clustering to minimize the maximum inter-cluster distance,” *Theoretical Computer Science, Elsevier Science B.V.*, vol. 38, pp. 293–306, 1985.

[55] P. Potikas, “The k-center problem - Approximation Algorithms.” Spring, 2009.

[56] P. Krishnan, D. Raz, and Y. Shavitt, “The Cache Location Problem,” *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 568–582, 2000.

[57] M. Yang and Z. Fei, "A Model for Replica Placement in Content Distribution Networks for Multimedia Applications," in *IEEE International Conference on Communications (ICC)*, vol. 1, pp. 557–561, 2003.

[58] X. Gu and C. Ding, "On the Theory and Potential of LRU-MRU Collaborative Cache Management," *ACM SIGPLAN Notices*, vol. 46, no. 11, pp. 43–54, 2011.

[59] E. Petrank and D. Rawitz, "The Hardness of Cache Conscious Data Placement," *ACM SIGPLAN Notices*, vol. 37, no. 1, pp. 101–112, 2002.

[60] S. Chaudhuri, N. Garg, and R. Ravi, "Generalized k-Center Problems," *Information Processing Letters*, 1996.

[61] S. Chaudhuri, N. Garg, and R. Ravi, "The P-neighbor K-center Problem," *Information Processing Letters*, vol. 65, no. 3, pp. 131–134, 1998.

[62] K. Jain and V. V. Vazirani, "Approximation Algorithms for Metric Facility Location and k-Median Problems Using the Primal-dual Schema and Lagrangian Relaxation," *Journal of the ACM (JACM)*, vol. 48, no. 2, pp. 274–296, 2001.

[63] J. F. Stollsteimer, *The effect of technical change and output expansion on the optimum number, size, and location of pear marketing facilities in a California pear producing region.* PhD thesis, University of California, Berkeley, California, 1961.

[64] M. L. Balinski, "Integer Programming: Methods, Uses, Computations," tech. rep., University of Pennsylvania and MATHEMATICA, 1965.

[65] J. Kleinberg and Éva Tardos, *Algorithms Design*. Pearson-Addison Wesley, 2013. Lecture slides by Kevin Wayne.

[66] Z. Drezner, ed., *Facility Location: A Survey of Applications and Methods*. Springer-Verlag New York, 1995.

[67] V. Verter, *Foundations of Location Analysis*, vol. 155, ch. 2 Uncapacitated and Capacitated Facility Location Problems, pp. 25–37. Boston, MA: Springer US, 2011.

[68] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences avec les Mémoires de Mathématique & de Physique, pour la méme Année (2e partie)*, pp. 666–704, 1781.

[69] R. C. Tryon, *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards Brothers, 1943.

[70] Rosetta Code, "K-means++ clustering." `https://rosettacode.org/wiki/K-means%2B%2B_clustering` (2017-02-14), 2017.

[71] S. Khuller and Y. J. Sussmann, "The Capacitated K-Center Problem," *SIAM Journal on Discrete Mathematics (SIDMA)*, vol. 13, no. 3, pp. 403–418, 2000.

[72] R. L. Francis, F. McGinnis, and J. A. White, *Facility Layout and Location: An Analytical Approach*, vol. 2. Pearson, 1991.

[73] W. Zhui, X. Wang, L. Yao, and Q. Fei, “Simulation run allocation for optimization of order-up-to level configuration in a complex logistics,” in *IEEE International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, pp. 46–50, 2010.

[74] M. S. Daskin, *Network and Discrete Location: Models, Algorithms, and Applications*. John Wiley & Sons, Inc., 2013.

[75] T. Xi-feng, M. Hai-jun, and L. Xu-hong, “Logistics facility location model based on reliability within the supply chain,” in *IEEE International Conference on Management of Innovation and Technology (ICMIT)*, pp. 1099–1103, 2008.

[76] K. Carling, M. Han, J. Hakansson, and P. Rebreyend, “An empirical test of the gravity p-median model,” HUI Working Paper 90, HUI Research, 2013.

[77] L. Snyder, “Facility Location Under Uncertainty: A Review,” *IIE Transactions*, vol. 38, no. 7, pp. 537–554, 2006.

[78] V. V. Vazirani, *Approximation Algorithms*. New York: Springer, 2001.

[79] D. P. Williamson and D. B. Shmoys, *The Design of Approximation Algorithms*. Cambridge University Press, 2011.

[80] R. Rana and D. Garg, “Heuristic Approaches for K-Center Problem,” in *IEEE Advance Computing Conference (IACC)*, 2009.

[81] W. Zeng, *GIS Based Facility Location Planning with Different Types of Consumers*. PhD thesis, Department of Earth & Atmospheric Sciences, University of Alberta, Edmonton, Alberta, Canada, 2007.

[82] X. W. Qin and F. Qin, “Hybrid Particle Swarm Optimization Algorithm for the Logistics Network Design Problem Under Concave Cost Function,” in *Proceedings of the International Conference on Computer Science and Software Engineering (CSSE)*, vol. 1, pp. 1182–1186, IEEE Computer Society, 2008.

[83] Y.-J. Jang, S.-Y. Jang, B.-M. Chang, and J. Park, “A Combined Model of Network Design and Production/Distribution Planning for a Supply Network,” *Computers & Industrial Engineering*, vol. 43, no. 1-2, pp. 263–281, 2002.

[84] K. Alicke, *Planung und Betrieb von Logistiknetzwerken: Unternehmensübergreifendes Supply Chain Management*. Springer-Verlag Berlin Heidelberg, 2005.

[85] R. Vahrenkamp and D. C. Mattfeld, *Logistiknetzwerke: Modelle für Standortwahl und Tourenplanung*, vol. 1. Gabler, 2007.

[86] K. Haase and M. Hoppe, “Transportnetzgestaltung für Paketdienstleister,” *Journal of Business Economics*, vol. 78, no. 9, pp. 857–874, 2007.

[87] E. L. Lawler, J. K. Lenstra, and A. H. G. R. Kan, *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley Interscience Series, 1985.

[88] G. Mayer, *Strategische Logistikplanung von Hub&Spoke-Systemen*. Produktion und Logistik, Deutscher Universitätsverlag, 2001.

[89] J. Owsinski, J. Stanczak, A. Barski, K. Sep, and P. Sapiecha, “Graph based approach to the minimum hub problem in transportation network,” in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 1641–1648, 2015.

[90] G. Carofiglio, M. Gallo, L. Muscariello, and D. Perino, “Modeling data transfer in content-centric networking,” in *International Teletraffic Congress (ITC)*, pp. 111–118, 2011.

[91] G. Carofiglio, L. Mekinda, and L. Muscariello, “LAC: Introducing Latency-Aware Caching in Information-Centric Networks - Technical Report,” *ACM Computing Research Repository (CoRR)*, 2015.

[92] S. Jamin, C. Jin, A. Kurc, D. Raz, and Y. Shavitt, “Constrained Mirror Placement on the Internet,” in *Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 1, pp. 31–40, 2001.

[93] L.-C. Chen and H.-A. Choi, “Approximation Algorithms for Data Distribution with Load Balancing of Web Servers,” in *IEEE International Conference on Cluster Computing*, pp. 274–281, 2001.

[94] Y. Inaba, Y. Tanigawa, and H. Tode, “Content Retrieval Method in Cooperation with CDN and ICN-based In-network Guidance over IP Network,” in *Proceedings of the Conference on Local Computer Networks (LCN)*, pp. 454–457, IEEE Computer Society, 2015.

[95] J. Sun, S. Gao, W. Yang, and F. Huang, “Replica Placement Algorithms in Content Distribution Networks,” in *International Symposium on Information Engineering and Electronic Commerce (IEEC)*, pp. 1–4, 2010.

[96] P. B. Danzig, R. S. Hall, R. S. Hall, and M. F. Schwartz, “A Case for Caching File Objects Inside Internetworks,” in *Conference Proceedings on Communications Architectures, Protocols and Applications*, SIGCOMM, pp. 239–248, ACM, 1993.

[97] S. Borst, V. Gupta, and A. Walid, “Distributed Caching Algorithms for Content Distribution Networks,” in *IEEE International Conference on Computer Communications (INFOCOM)*, pp. 1–9, 2010.

[98] C. R. da Cunha, “Trace Analysis and its Applications to Performance Enhancements of Distributed Information Systems,” 1997.

[99] M. I. Andreica and N. Tapus, “Central Placement of Storage Servers in tree-like Content Delivery Networks,” in *IEEE Region 8 EUROCON (Europe, Middle East and Africa)*, pp. 1901–1908, 2009.

[100] G. Yao, J. Bi, Y. Li, and L. Guo, “On the Capacitated Controller Placement Problem in Software Defined Networks,” *IEEE Communications Letters*, vol. 18, no. 8, pp. 1339–1342, 2014.

[101] G. Carofiglio, V. Gehlen, and D. Perino, “Experimental Evaluation of Memory Management in Content-Centric Networking,” in *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2011.

[102] I. Psaras, R. G. Clegg, R. Landa, W. K. Chai, and G. Pavlou, “Modelling and Evaluation of CCN-caching Trees,” in *Proceedings of the International IFIP TC 6 Conference on Networking*, vol. 1 of *NETWORKING*, pp. 78–91, Springer-Verlag, 2011.

[103] J. Li, H. Wu, B. Liu, J. Lu, Y. Wang, X. Wang, Y. Zhang, and L. Dong, “Popularity-driven Coordinated Caching in Named Data Networking,” in *Proceedings of the Eighth Symposium on Architectures for Networking and Communications Systems (ANCS)*, pp. 15–26, ACM/IEEE, 2012.

[104] A. Dan and D. Towsley, “An Approximate Analysis of the LRU and FIFO Buffer Replacement Schemes,” in *Proceedings of the ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS, pp. 143–152, ACM, 1990.

[105] A. Chankhunthod, P. B. Danzig, C. Neerdaels, M. F. Schwartz, and K. J. Worrell, “A Hierarchical Internet Object Cache,” in *Proceedings of the Annual Conference on USENIX Annual Technical Conference*, ATEC, pp. 1–13, USENIX Association, 1996.

[106] Z. Wang, H. Jiang, Y. Sun, J. Li, J. Liu, and E. Dutkiewicz, “A K-coordinated Decentralized Replica Placement Algorithm for the Ring-based CDN-P2P Architecture,” in *Proceedings of the The IEEE Symposium on Computers and Communications (ISCC)*, pp. 811–816, IEEE Computer Society, 2010.

[107] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, “Measurement, Modeling, and Analysis of a Peer-to-peer File-sharing Workload,” *Operating Systems Review (SIGOPS)*, vol. 37, no. 5, pp. 314–329, 2003.

[108] M. Hefeeda and O. Saleh, “Traffic Modeling and Proportional Partial Caching for Peer-to-peer Systems,” *IEEE/ACM Transaction Networks*, vol. 16, no. 6, pp. 1447–1460, 2008.

[109] J. Zhou, X. Zhang, L. Bhuyan, and B. Liu, “Clustered K-Center: Effective Replica Placement in Peer-to-Peer Systems,” in *IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 2008–2013, 2007.

[110] F. L. Presti, N. Bartolini, and C. Petrioli, “Dynamic replica placement and user request redirection in content delivery networks,” in *IEEE International Conference on Communications (ICC)*, vol. 3, pp. 1495–1501, 2005.

[111] S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang, “On the Placement of Internet Instrumentation,” in *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 1, pp. 295–304, 2000.

[112] P. Radoslavov, R. Govindan, and D. Estrin, “Topology-informed Internet Replica Placement,” *Computer Communications*, vol. 25, no. 4, pp. 384–392, 2002.

[113] X. Li and J. Li, *Quality-Based Content Delivery over the Internet*, vol. 1. Shanghai Jiao Tong University Press: Shanghai and Springer-Verlag Berlin Heidelberg, 2011.

[114] R. Buyya, M. Pathan, and A. Vakali, *Content Delivery Networks*, vol. 9. Springer-Verlag Berlin Heidelberg, 2008.

[115] E. Nygren, R. K. Sitaraman, and J. Sun, "The Akamai Network: A Platform for High-performance Internet Applications," *ACM Special Interest Group on Operating Systems (SIGOPS)*, vol. 44, no. 3, pp. 2–19, 2010.

[116] Forrester Consulting, "An Updated Look At Consumer Reaction To A Poor Online Shopping Experience," *eCommerce Web Site Performance Today*, 2009.

[117] S. Androutsellis-Theotokis and D. Spinellis, "A Survey of Peer-to-peer Content Distribution Technologies," *ACM Computing Surveys*, vol. 36, no. 4, pp. 335–371, 2004.

[118] A. Sawall, "Telekom und CDN-Anbieter Akamai arbeiten zusammen," *Golem Media GmbH*, 2016. `http://www.golem.de/news/content-delivery-network-telekom-und-cdn-anbieter-akamai-arbeiten-zusammen-1603-119775.html` (2017-02-14).

[119] A. Dahlmann, *Replica Placement in Content Delivery Networks: Model and Methods*. PhD thesis, Technische Universität München, 2011.

[120] O. Ercetin, *Market- Based Resource Allocation for Content Delivery in the Internet*. PhD thesis, Faculty of the Graduate School of the University of Maryland, 2002.

[121] I. Baev, R. Rajaraman, and C. Swamy, "Approximation Algorithms for Data Placement Problems," *SIAM Journal on Computing*, vol. 38, no. 4, pp. 1411–1429, 2008.

[122] J. Coppens, T. Wauters, F. de Turck, B. Dhoedt, and P. Demeester, "Design and Performance of a Self-Organizing Adaptive Content Distribution Network," in *IEEE/IFIP Network Operations and Management Symposium (NOMS)*, pp. 534–545, 2006.

[123] G. Pallis and A. Vakali, "Insight and Perspectives for Content Delivery Networks," *Communications of the ACM*, vol. 49, no. 1, pp. 101–106, 2006.

[124] G. Pallis, A. Vakali, K. Stamos, A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos, "A latency-based object placement approach in content distribution networks," in *Third Latin American Web Congress (LA-WEB)*, 2005.

[125] R. Mijumbi, *Self-managed Resources in Network Virtualisation Environments*. PhD thesis, Universitat Politècnica de Catalunya, 2014.

[126] J. P. Mulerikkal, "An Architecture for Distributed Content Delivery Network," Master's thesis, Computer Science and Information Technology, Melbourne, Victoria, Australia, 2007.

[127] L. Qiu, V. N. Padmanabhan, and G. M. Voelker, "On the Placement of Web Server Replicas," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*, 2001.

[128] B. Frank, I. Poese, Y. Lin, G. Smaragdakis, A. Feldmann, B. Maggs, J. Rake, S. Uhlig, and R. Weber, "Pushing CDN-ISP Collaboration to the Limit," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 3, pp. 34–44, 2013.

[129] N. Herbaut, D. Négru, Y. Chen, P. A. Frangoudis, and A. Ksentini, "Content Delivery Networks as a Virtual Network Function: a win-win ISP-CDN collaboration," in *IEEE Global Communications Conference*, 2016.

[130] R. Alimi, R. Penno, Y. Yang, S. Kiesel, S. Previdi, W. Roome, S. Shalunov, and R. Woundy, "Application-Layer Traffic Optimization (ALTO) Protocol - Request for Comments: 7285," standards track, Internet Engineering Task Force (IETF), 2014.

[131] R. Krishnan, H. V. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao, "Moving Beyond End-to-end Path Information to Optimize CDN Performance," in *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference*, pp. 190–201, 2009.

[132] D. Tuncer, M. Charalambides, R. Landa, and G. Pavlou, "More Control Over Network Resources: An ISP Caching Perspective," in *IEEE/IFIP International Conference on Network and Service Management (CNSM)*, 2013.

[133] I. Contreras and E. Fernández, "General network design: A unified view of combined location and network design problems," *European Journal of Operational Research*, vol. 219, no. 3, pp. 680–697, 2012.

[134] H. Hu, Y. Wen, T. S. Chua, Z. Wang, J. Huang, W. Zhu, and D. Wu, "Community based effective social video contents placement in cloud centric CDN network," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2014.

[135] F. Wang, J. Liu, and M. Chen, "CALMS: Cloud-assisted live media streaming for globalized demands with time/region diversities," in *IEEE International Conference on Computer Communications (INFOCOM)*, pp. 199–207, 2012.

[136] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You Are Who You Know: Inferring User Profiles in Online Social Networks," in *International Conference on Web Search and Data Mining (WSDM)*, pp. 251–260, ACM, 2010.

[137] W. Jiang, R. Zhang-Shen, J. Rexford, and M. Chiang, "Cooperative Content Distribution and Traffic Engineering in an ISP Network," in *Proceedings of the International Joint Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS, pp. 239–250, ACM, 2009.

[138] Z. Xu and L. Bhuyan, "QoS-aware object replica placement in CDNs," in *IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 2, pp. 862–866, 2005.

[139] D. Papadimitriou, D. Colle, and P. Demeester, "Mixed-integer optimization for the combined capacitated facility location-routing problem," in *International Conference on the Design of Reliable Communication Networks (DRCN)*, pp. 14–22, 2016.

[140] N. Kamiyama, "CCN Economics: Can ISPs Profit by Introducing CCN?," in *International Conference on Network and Service Management (CNSM) and Workshop*, pp. 242–247, 2014.

[141] W. K. Chai, D. He, I. Psaras, and G. Pavlou, "Cache 'Less for More' in Information-centric Networks," in *Proceedings of the International IFIP TC 6 Conference on Networking*, vol. 1, pp. 27–40, Springer-Verlag, 2012.

[142] I. Castro, R. Stanojevic, and S. Gorinsky, "Using Tuangou to Reduce IP Transit Costs," *IEEE/ACM Transaction Networks*, vol. 22, no. 5, pp. 1415–1428, 2014.

[143] K. Fischer, *Standortplanung unter Berücksichtigung verschiedener Marktbedingungen*. Physica-Schriften zur Betriebswirtschaft, 1997.

[144] M. Eßig, E. Hoffmann, and W. Stölzle, *Supply Chain Management*. Franz Vahlen München, 2013.

[145] A. J. Abu and B. Bensaou, "Modelling the Pending Interest Table occupancy in CCN with interest timeout and retransmission," in *IEEE Conference on Local Computer Networks (LCN)*, pp. 37–45, 2015.

[146] S. Tarnoi, V. Suppakitpaisarn, W. Kumwilaisak, and Y. Ji, "Performance Analysis of Probabilistic Caching Scheme Using Markov Chains," in *Proceedings of the Conference on Local Computer Networks (LCN)*, pp. 46–54, IEEE Computer Society, 2015.

[147] I. Psaras, W. K. Chai, and G. Pavlou, "Probabilistic In-network Caching for Information-centric Networks," in *Proceedings of the Workshop on Information-centric Networking (ICN)*, pp. 55–60, ACM, 2012.

[148] Z. Feng, M. Xu, Y. Yang, Q. Li, Y. Wang, Q. Li, B. Ohlman, and M. Chen, "Joint optimization of content replication and Traffic Engineering in ICN," in *IEEE Conference on Local Computer Networks (LCN)*, pp. 55–63, 2015.

[149] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 1, pp. 126–134, 1999.

[150] V. Almeida, A. Bestavros, M. Crovella, and A. deOliveira, "Characterizing Reference Locality in the WWW," tech. rep., Boston University, MA, USA, 1996.

[151] V. F. Almeida, M. G. Cesário, R. C. Fonseca, W. M. Jr., and C. D. Murta, "Analyzing the Behavior of a Proxy Server in Light of Regional and Cultural Issues," *Proceedings of the International Web Content Caching and Distribution Workshop (WCW)*, 1998.

[152] C. Cunha, A. Bestavros, and M. Crovella, "Characteristics of WWW Client-based Traces," tech. rep., Boston University, MA, USA, 1995.

[153] S. Glassman, "A Caching Relay for the World Wide Web," in *Selected Papers of the Conference on World-Wide Web*, pp. 165–173, Elsevier Science Publishers B. V., 1994.

[154] N. Nishikawa, T. Hosokawa, Y. Mori, K. Yoshida, and H. Tsuji, "Memory-based Architecture for Distributed WWW Caching Proxy," in *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 205–214, Elsevier Science Publishers B. V., 1998.

[155] L. A. Adamic and B. A. Huberman, “Zipf’s law and the Internet,” *Glottometrics 3*, pp. 143–150, 2002. `http://www.hpl.hp.com/research/idl/papers/ranking/adamicglottometrics.pdf` (2017-02-14).

[156] C. Anastasiades, A. Gomes, R. Gadow, and T. Braun, “Persistent caching in information-centric networks,” in *IEEE Conference on Local Computer Networks (LCN)*, pp. 64–72, 2015.

[157] P. Maillé, K. Pires, G. Simon, and B. Tuffin, “How Neutral is a CDN? An economic approach,” in *International Conference on Network and Service Management (CNSM) and Workshop*, pp. 336–339, 2014.

[158] M. Zink, K. Suh, Y. Gu, and J. Kurose, “Characteristics of YouTube Network Traffic at a Campus Network - Measurements, Models, and Implications,” *Journal Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 53, no. 4, pp. 501–514, 2009.

[159] S. A. Chowdhury and D. J. Makaroff, “Popularity Growth Patterns of YouTube Videos: A Category-based Study,” in *International Conference on Web Information Systems and Technologies (WEBIST)*, 2013.

[160] M. P. Wittie, V. Pejovic, L. Deek, K. C. Almeroth, and B. Y. Zhao, “Exploiting Locality of Interest in Online Social Networks,” in *Proceedings of the International Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, pp. 25:1–25:12, ACM, 2010.

[161] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft, “Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades,” in *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 457–466, ACM, 2011.

[162] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida, “On Word-of-mouth Based Discovery of the Web,” in *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference (IMC)*, pp. 381–396, 2011.

[163] X. Wang, M. Chen, Z. Han, D. O. Wu, and T. T. Kwon, “TOSS: Traffic offloading by social network service-based opportunistic sharing in mobile social networks,” in *IEEE Conference on Computer Communications (INFOCOM)*, pp. 2346–2354, 2014.

[164] M. D. Choudhury, H. Sundaram, A. John, D. D. Seligmann, and A. Kelliher, “”Birds of a Feather”: Does User Homophily Impact Information Diffusion in Social Media?,” *Cornell University Library, Computers and Society*, 2010.

[165] C. Fricker, P. Robert, J. Roberts, and N. Sbihi, “Impact of Traffic Mix on Caching Performance in a Content-Centric Network,” *Cornell University Library, Computers and Society*, 2012.

[166] S. Spagna, M. Liebsch, R. Baldessari, S. Niccolini, S. Schmid, R. Garroppo, K. Ozawa, and J. Awano, “Design Principle of an Operator-Owned HiHigh Distributed Content Delivery Network,” *IEEE Communications Magazine*, 2013.

[167] M. Claeys, D. Tuncer, J. Famaey, M. Charalambides, S. Latré, F. D. Turck, and G. Pavlou, "Towards multi-tenant cache management for ISP networks," in *European Conference on Networks and Communications (EuCNC)*, pp. 1–5, 2014.

[168] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," *Communications of the ACM*, vol. 43, no. 8, pp. 142–151, 2000.

[169] Reginale Wertschöpfungspartnerschaften, "Erfolgsfaktoren," *nova-Institut GmbH*, o. J. `http://www.regionale-wertschoepfung.info/` (2017-02-14).

[170] E. H. Frazelle, *Supply Chain Strategy: The Logistics of Supply Chain Management*, vol. 1. McGraw-Hill, 2001.

[171] FLAMINGO Consortium ICT-318488, "WP6 - Automated Configuration and Repair, Deliverable D6.1 - First Year Report on Automated Configuration and Repair," *European Seventh Framework Network of Excellence*, 2013.

[172] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer Publishing Company, Incorporated, 2008.

[173] M. Dyer and A. Frieze, "A Simple Heuristic for the P-centre Problem," *Operations Research Letters*, vol. 3, no. 6, pp. 285–288, 1985.

[174] T. Feder and D. Greene, "Optimal Algorithms for Approximate Clustering," in *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pp. 434–444, 1988.

[175] S. Dasgupta, "Clustering in metric spaces." `http://cseweb.ucsd.edu/~dasgupta/291-geom/kcenter.pdf` (2017-02-14), 2013.

[176] D. S. Hochbaum and D. B. Shmoys, "A Best Possible Heuristic for the k-Center Problem," in *Mathematics of Operations Research*, vol. 10 (2), pp. 180–184, INFORMS, 1985.

[177] J. Plesnik, "A heuristic for the p-center problems in graphs," *Discrete Applied Mathematics*, vol. 17, no. 3, pp. 263–268, 1987.

[178] T. A. Feo and M. G. C. Resende, "Greedy Randomized Adaptive Search Procedures," Journal of Global Optimization 6, Kluwer Academic Publishers, 1995.

[179] J. Hart and A. W. Shogan, "Semi-greedy Heuristics: An Empirical Study," *Operations Research Letters*, vol. 6, no. 3, pp. 107–114, 1987.

[180] M. G. Resende and C. C. Ribeiro, "Greedy randomized adaptive search procedures," *AT&T Labs Research Technical Report*, vol. 2, 2002.

[181] S. Z. Selim and M. A. Ismail, "K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 1, pp. 81–87, 1984.

[182] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[183] Q. Du, V. Faber, and M. Gunzburger, "Centroidal Voronoi Tessellations: Applications and Algorithms," *Society for Industrial and Applied Mathematics (SIAM) Review*, vol. 41, no. 4, pp. 637–676, 1999.

[184] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[185] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[186] S. Dasgupta, "Algorithms for k-means clustering." `http://cseweb.ucsd.edu/~dasgupta/291-geom/kmeans.pdf` (2017-02-14), 2013.

[187] S. H. Ganesh and C. Chandrasekar, "A novel priority based data mining algorithm using improved K-means clustering for detecting protein sequence from dataset," in *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–4, 2010.

[188] L. Morissette and S. Chartier, "The k-means clustering technique: General considerations and implementation in Mathematica," *Tutorials in Quantitative Methods for Psychology*, vol. 9, no. 1, pp. 15–24, 2013.

[189] A. Choudhary, "Survey on K-Means and Its Variants," *International Journal of Innovative Research in Computerand Communication Engineering*, vol. 4, no. 1, 2016.

[190] J. P. Pedroso, *Optimization with Gurobi and Python*. INESC Porto and Universidade do Porto„ 1 ed., 2011. `http://www.dcc.fc.up.pt/~jpp/seminars/azores/gurobi-intro.pdf` (2017-02-14).

[191] S. Dasgupta, "The k-median clustering problem." `http://cseweb.ucsd.edu/~dasgupta/291-geom/kmedian.pdf` (2016-05-20), 2013.

[192] I. Gurobi Optimization, "Gurobi Optimizer Reference Manual," 2015. `http://www.gurobi.com` (2017-02-14).

[193] "IBM ILOG CPLEX Optimizer," 2010. `http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/` (2017-02-14).

[194] G. B. Dantzig, *Linear Programming and Extensions*. The RAND Cooperation, Princeton: University Press, 1963.

[195] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoys, "A Constant-factor Approximation Algorithm for the K-median Problem," in *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pp. 1–10, 1999.

[196] S. H. Owen and M. S. Daskin, "Strategic facility location: A review," *European Journal of Operational Research*, vol. 111, pp. 423–447, 1998.

[197] K. Jain and V. V. Vazirani, “Primal-Dual Approximation Algorithms for Metric Facility Location and k-Median Problems,” in *Journal of the ACM (JACM)*, pp. 2–13, 1999.

[198] J. Bisschop, *AIMMS - Optimization Modeling*. Lulu: Online Self Publishing Book & eBook Company, 2006.

[199] M. X. Goemans and D. P. Williamson, “The Primal-dual Method for Approximation Algorithms and Its Application to Network Design Problems,” in *Approximation Algorithms for NP-Hard Problems* (D. S. Hochbaum, ed.), pp. 144–191, Boston, MA, USA: PWS Publishing Co., 1997.

[200] A. Archer, R. Rajagopalan, and D. B. Shmoys, “Lagrangian relaxation for the k-median problem: new insights and continuity properties.,” in *Proceedings of the Annual European Symposium on Algorithms*, vol. 11, pp. 31–42, Springer Science & Business Media, 2003.

[201] S. Chawla, “CS 787: Advanced Algorithms - Lecture 10,” 2009. `http://pages.cs.wisc.edu/~shuchi/courses/787-F09/scribe-notes/lec10.pdf` (2017-02-14).

[202] M. Mahdian, Y. Ye, and J. Zhang, “Improved Approximation Algorithms for Metric Facility Location Problems,” in *Proceedings of the International Workshop on Approximation Algorithms for Combinatorial Optimization (APPROX)*, pp. 229–242, Springer-Verlag, 2002.

[203] A. A. Ageev and M. Sviridenko, “An 0.828-approximation Algorithm for the Uncapacitated Facility Location Problem,” *Discrete Applied Mathematics*, vol. 93, pp. 149–156, 1999.

[204] K. Jain and V. V. Vazirani, “An Approximation Algorithm for the Fault Tolerant Metric FacilityLocation Problem,” *Algorithmica*, vol. 38, no. 3, pp. 433–439, 2004.

[205] R. E. Gomory, “Solving Linear Programming Problems in Integers,” *Symposia in Applied Mathematics*, vol. 10, pp. 1–6, 1960.

[206] M. Charikar and S. Li, “A Dependent LP-rounding Approach for the k-Median Problem,” in *Proceedings of the international colloquium conference on Automata, Languages, and Programming (ICALP)* (A. Czumaj, K. Mehlhorn, A. Pitts, and R. Wattenhofer, eds.), vol. 1, pp. 194–205, Springer Berlin Heidelberg, 2012.

[207] M. Cygan, M. T. Hajiaghayiy, and S. Khullery, “LP Rounding for k-Centers with Non-uniform Hard Capacities,” in *Proceedings of the IEEE Annual Symposium on Foundations of Computer Science (FCOS)*, pp. 273–282, 2012.

[208] E. Schubert, A. Koos, T. Emrich, A. Züfle, K. A. Schmid, and A. Zimek, “A Framework for Clustering Uncertain Data,” *Proceedings of the VLDB Endowment (PVLDB)*, vol. 8, no. 12, pp. 1976–1979, 2015. `http://www.vldb.org/pvldb/vol8/p1976-schubert.pdf` (2017-02-14).

[209] L. Kaufman and P. J. Rousseeuw, *Partitioning Around Medoids (Program PAM)*, pp. 68–125. John Wiley & Sons, Inc., 2008.

[210] E. Schubert, “Invoking ELKI from Java,” 2016. http://elki.dbs.ifi.lmu.de/wiki/HowTo/InvokingELKIFromJava (2017-02-14).

[211] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[212] T. Uhlig, “Serein - a Framework to Model Metaheuristics,” 2013. http://sourceforge.net/projects/serein/ (2017-02-14).

[213] T. Uhlig, *Self-Replicating Individuals*. Informatik, Dissertation, Universität der Bundeswehr München: Verlag Dr. Hut, 2015.

[214] J. H. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

[215] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by Simulated Annealing,” *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[216] D. Hadka, “MOEA Framework - A Free and Open Source Java Framework for Multiobjective Optimization,” tech. rep., GNU Lesser General Public License, 2011. http://moeaframework.org/ (2017-02-14).

[217] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II,” *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[218] A. J. Nebro, J. J. Durillo, F. Luna, B. Dorronsoro, and E. Alba, “MOCell: A Cellular Genetic Algorithm for Multiobjective Optimization,” *International Journal of Intelligent Systems*, vol. 24, pp. 726–746, July 2009.

[219] J. Knowles and D. Corne, “The Pareto archived evolution strategy: a new baseline algorithm for Pareto multiobjective optimisation,” in *Congress on Evolutionary Computation (CEC)*, vol. 1, pp. 105–112, 1999.

[220] E. Zitzler, M. Laumanns, and L. Thiele, “SPEA2: Improving the Strength Pareto Evolutionary Algorithm,” tech. rep., Swiss Federal Institute of Technology (ETH) Zurich, 2001.

[221] J. D. Schaffer, “Multiple Objective Optimization with Vector Evaluated Genetic Algorithms,” in *International Conference on Genetic Algorithms*, (Hillsdale, NJ, USA), pp. 93–100, L. Erlbaum Associates Inc., 1985.

[222] J. Durillo, A. Nebro, and E. Alba, “The jMetal Framework for Multi-Objective Optimization: Design and Architecture,” in *Congress on Evolutionary Computation (CEC) 2010*, pp. 4138–4325, 2010.

[223] S. Bleuler, M. Laumanns, L. Thiele, and E. Zitzler, “PISA - A Platform and Programming Language Independent Interface for Search Algorithms,” in *International Conference on Evolutionary Multi-criterion Optimization (EMO)*, pp. 494–508, Springer, 2003.

[224] DIN 66001:1983-12, *Informationsverarbeitung, Sinnbilder und ihre Anwendung*, 1983.

[225] J. M. Kim, J. Choi, J. Kim, S. H. Noh, S. L. Min, Y. Cho, and C. S. Kim, "A Low-Overhead, High Performance Unified Buffer Management Scheme that Exploits Sequential and Looping References," *Proceedings of the USENIX Symposium on Operating System Design and Implementation*, vol. 4, pp. 119–134, 2000.

[226] M. Baker, J. Hartman, M. Kupfer, K. Shirriff, and J. Ousterhout, *Measurements of a Distributed File System*, pp. 198–212. SOSP, New York, NY, USA: ACM, 1991.

[227] G. K. Zipf, "Selected Studies of the Principle of Relative Frequency in Language," *Harvard University Press*, 1932.

[228] G. K. Zipf, "Human Behavior and the Principle of Least Effort," *Addison-Wesley*, 1949.

[229] M. West, "The mystery of zipf," tech. rep., Plus Magazine, 2008. `https://plus.maths.org/content/os/latestnews/may-aug08/Zipf/index` (2017-02-14).

[230] J. W. von Goethe, *Faust. Der Tragödie erster Teil*. Die freie digitale Bibliothek, 1808. `http://www.digbib.org/Johann_Wolfgang_von_Goethe_1749/Faust_I` (2017-02-14).

[231] G. Maier, A. Feldmann, V. Paxson, and M. Allman, "On Dominant Characteristics of Residential Broadband Internet Traffic," in *ACM SIGCOMM Conference on Internet Measurement (IMC)*, pp. 90–102, 2009.

[232] J. Nielsen, "Traffic-Log Muster." The swiss usability center, 2006. `http://www.usability.ch/news/traffic-log-muster.html` (2017-02-14).

[233] B. B. Mandelbrot, "An information theory of the statistical structure of language," in *Communication Theory* (W. Jackson, ed.), pp. 503–512, New York: Academic Press, 1953.

[234] A. Fischer, J. F. Botero, M. T. Beck, H. de Meer, and X. Hesselbach, "Virtual Network Embedding: A Survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1888–1906, 2013.

[235] A. Sidiropoulos, G. Pallis, D. Katsaros, K. Stamos, A. Vakali, and Y. Manolopoulos, "Prefetching in Content Distribution Networks via Web Communities Identification and Outsourcing," *World Wide Web*, vol. 11, no. 1, pp. 39–70, 2008.

[236] C. Royer, *Simultane Optimierung von Produktionsstandorten, Produktionsmengen und Distributionsgebieten*. Herbert Utz Verlag, 2001.

[237] W. M. Mularie, "Department of Defense: World Geodetic System 1984," tech. rep., National Imagery and Mapping Agency, 2000.

[238] Google Inc., "Google Maps," 2017. `https://www.google.de/maps` (2016-02-14).

[239] E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs," *Numerische Mathematik*, 1959.

[240] R. W. Floyd, “Algorithm 97 (shortest path),” in *Communications of the ACM*, no. 5 in CACM, pp. 345–351, 1962.

[241] S. Warshall, “A theorem on boolean matrices,” in *Journal of the ACM*, no. 1 in JACM, pp. 11–12, 1962.

[242] P. E. Hart, N. J. Nilsson, and B. Raphael, “A Formal Basis for the Heuristic Determination of Minimum Cost Paths,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[243] T. P. Kirkman, “On the representation of polyhedra,” *Philosophical Transactions of the Royal Society of London Series A 146*, pp. 413–418, 1856.

[244] OpenStreetMap contributors, “Planet dump,” 2015. `http://planet.openstreetmap.org` (2016-02-14).

[245] D. Z. Chen, J. Li, H. Liang, and H. Wang, “Matroid and Knapsack Center Problems,” *Computing Research Repository (CoRR)*, 2013.

[246] H.-P. Kriegel, P. Kröger, and A. Zimek, “Clustering High-dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, pp. 1–58, 2009.

[247] L. Parsons, E. Haque, and H. Liu, “Subspace Clustering for High Dimensional Data: A Review,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.

[248] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, vol. 8. Iowa State University Press, 8 ed., 1989.

[249] Royal Society (Great Britain), *Proceedings of the Royal Society of London*, vol. 58. University of Michigan: Taylor & Francis, 1895.

[250] M. G. Kendall, “A New Measure of Rank Correlation,” *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[251] C. Spearman, “The Proof and Measurement of Association between Two Things,” *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[252] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, “A Simulated Annealing-Based Multiobjective Optimization Algorithm: AMOSA,” in *Proceedings of the IEEE Transactions on Evolutionary Computation*, vol. 12-3, pp. 269–283, 2008.

[253] T. Drezner and Z. Drezner, “The gravity multiple server location problem,” *Computers & Operations Research*, vol. 38, no. 3, pp. 694 – 701, 2011.

[254] O. Grün, *Industriebetriebslehre: Das Wirtschaften in Industrieunternehmen; Kapitel 6 - Industrielle Materialwirtschaft*, vol. 2. Franz Vahlen München, 1994.

[255] W. Dangelmaier and H.-J. Warnecke, *Fertigungslenkung: Planung und Steuerung des Ablaufs der diskreten Fertigung*. Springer-Verlag Berlin Heidelberg, 1997.

[256] K. Schade, *Stochastische Optimierung: Bestandsoptimierung in mehrstufigen Lagernetzwerken*. Vieweg Teubner Verlag, 2012.

[257] T. Gudehus, *Dynamische Disposition: Strategien, Algorithmen und Werkzeuge zur optimalen Auftrags-, Bestands- und Fertigungsdisposition*. Springer-Verlag Berlin Heidelberg, 2012.

[258] R. Manzini, *Warehousing in the Global Supply Chain: Advanced Models*. Springer-Verlag London, 2012.

[259] P. M. Krobath, *Die kostenoptimale Bestellpolitik für die Beschaffung von legiertem Halbzeug aus Stahl*. PhD thesis, Montanuniversität Leoben, 2010.

[260] M. Karlsson and C. Karamanolis, "Choosing Replica Placement Heuristics for Wide-Area Systems," in *Proceedings of the International Conference on Distributed Computing Systems (ICDCS)*, pp. 350–359, IEEE Computer Society, 2004.

[261] L. A. Bélády, "A Study of Replacement Algorithms for a Virtual-storage Computer," *IBM Systems Journal*, vol. 5, no. 2, pp. 78–101, 1966.

[262] R. L. Kruse, *Data Structures and Program Design*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1984.

[263] F. J. Corbato, "A Paging Experiment with the Multics System," *Massachusetts Institute of Technology Cambridge Project MAC*, p. 20, 1968.

[264] A. S. Tanenbaum, *Modern Operating Systems*. Upper Saddle River, NJ, USA: Prentice Hall Press, 4th ed., 2007.

[265] E. J. O'Neil, P. E. O'Neil, and G. Weikum, "The LRU-K Page Replacement Algorithm for Database Disk Buffering," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 297–306, 1993.

[266] D. Lee, J. Choi, J. H. Kim, S. H. Noh, S. L. Min, Y. Cho, and C. S. Kim, "LRFU: A Spectrum of Policies That Subsumes the Least Recently Used and Least Frequently Used Policies," *IEEE Transactions on Computers*, vol. 50, no. 12, pp. 1352–1361, 2001.

[267] J. T. Robinson and M. V. Devarakonda, "Data Cache Management Using Frequency-based Replacement," in *Proceedings of the Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS, pp. 134–142, ACM, 1990.

[268] Y. Zhou, Z. Chen, and K. Li, "Second-Level Buffer Cache Management," *IEEE Transactions on Parallel and Distributed Systems*, vol. 15, no. 6, pp. 505–519, 2004.

[269] N. Megiddo and D. S. Modha, "Outperforming LRU with an Adaptive Replacement Cache Algorithm," *Computer*, vol. 37, no. 4, pp. 58–65, 2004.

[270] S. Bansal and D. S. Modha, "CAR: Clock with Adaptive Replacement," in *Proceedings of the 3rd USENIX Conference on File and Storage Technologies*, FAST, pp. 187–200, USENIX Association, 2004.

[271] S. Jiang and X. Zhang, “LIRS: An Efficient Low Inter-reference Recency Set Replacement Policy to Improve Buffer Cache Performance,” in *Proceedings of the International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS, pp. 31–42, ACM, 2002.

[272] R. B. Gramacy, M. K. Warmuth, S. A. Brandt, and I. Ari, “Adaptive Caching by Refetching,” in *Advances in Neural Information Processing Systems 15 (NIPS)*, pp. 1465–1472, 2002.

[273] Y. Zhou, J. Philbin, and K. Li, “The Multi-Queue Replacement Algorithm for Second Level Buffer Caches,” in *Proceedings of the General Track: USENIX Annual Technical Conference*, pp. 91–104, USENIX Association, 2001.

[274] L. A. Bélády, R. A. Nelson, and G. S. Shedler, “An Anomaly in Space-time Characteristics of Certain Programs Running in a Paging Machine,” *Communications of the ACM*, vol. 12, no. 6, pp. 349–353, 1969.

[275] P. W. Berg, S. Isaacs, and L. Blodgett, *Airborne Fulfillment Center Utilizing Unmanned Aerial Vehicles For Item Delivery*, 2016. Amazon Technologies, Inc., Patent US 9,305280 B1.

[276] A. Borshchev, “AnyLogic 7: New Release Presentation,” in *Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, WSC, (Piscataway, NJ, USA), pp. 4106–4106, IEEE Press, 2013. `http://www.anylogic.de` (2016-02-14).

[277] Dial Telecom, “Backbone network,” 2014. `http://www.dialtelecom.cz/en/optical-network/backbone-network/` (2017-02-14).

[278] V. N. Padmanabhan and L. Subramanian, “An Investigation of Geographic Mapping Techniques for Internet Hosts,” in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM, pp. 173–185, ACM, 2001.

[279] A. Dahnert, “HawkEyes: An advanced IP Geolocation approach: IP Geolocation using semantic and measurement based techniques,” in *Worldwide Cybersecurity Summit (WCS)*, 2011.

[280] S. Laki, P. Matray, P. Haga, T. Sebok, I. Csabai, and G. Vattay, “Spotter: A model based active geolocation service,” in *INFOCOM Conference*, pp. 3173–3181, Institute of Electrical and Electronics Engineers, Inc., 2011.

[281] B. Eriksson, P. Barford, B. Maggs, and R. Nowak, “Posit: An Adaptive Framework for Lightweight IP Geolocation,” tech. rep., BU/CS, July 2011.

[282] J. S. Otto, M. A. Sánchez, J. P. Rula, and F. E. Bustamante, “Content Delivery and the Natural Evolution of DNS: Remote Dns Trends, Performance Issues and Alternative Solutions,” in *Internet Measurement Conference (IMC)*, pp. 523–536, ACM, 2012.

[283] C. Contavalli, W. van der Gaast, S. Leach, D. Lawrence, and W. Kumari, “Client Subnet in DNS Requests draft-vandergaast-dnsop-edns-client-subnet-00.” IETF Trust: dnsop, 2015.

[284] B. Wong, I. Stoyanov, and E. G. Sirer, “Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts,” in *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, pp. 23–23, USENIX Association, 2007.

[285] R. Koch, M. Golling, L. Stiemert, and G. D. Rodosek, “Using Geolocation for the Strategic Preincident Preparation of an IT Forensics Analysis,” *IEEE Systems Journal*, vol. 10, no. 4, pp. 1338–1349, 2016.

[286] P. Endo and D. Sadok, “Whois Based Geolocation: A Strategy to Geolocate Internet Hosts,” in *IEEE International Conference on Advanced Information Networking and Applications (AINA)*, pp. 408–413, 2010.

[287] B. Huffaker, M. Fomenkov, and k. claffy, “Geocompare: a comparison of public and commercial geolocation databases - Technical Report,” tech. rep., Cooperative Association for Internet Data Analysis (CAIDA), 2011.

[288] C. Davis, P. Vixie, T. Goodwin, and A. Dickinson, “A means for expressing location information in the domain name system.” Network Working Group - Request for Comments: 1876, 1996.

[289] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, “Measuring ISP Topologies with Rocketfuel,” *IEEE/ACM Trans. Netw.*, vol. 12, 2004.

[290] Whois.com, “IP Lookup Tool,” o. J. `https://www.whois.com/` (2017-02-14).

[291] I. MaxMind, “MaxMind Geolocation Service,” o. J. `http:www.maxmind.com/` (2017-02-14).

[292] D. Moore, R. Periakaruppan, J. Donohoe, and K. Claffy, “Where in the world is netgeo.caida.org?,” tech. rep., INET, 2000.

[293] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang, “Mining the Web and the Internet for Accurate IP Address Geolocations,” in *INFOCOM Conference*, pp. 2841–2845, Institute of Electrical and Electronics Engineers, Inc., 2009.

[294] S. S. Siwpersad, B. Gueye, and S. Uhlig, “Assessing the Geographic Resolution of Exhaustive Tabulation for Geolocating Internet Hosts,” in *Proceedings of the International Conference on Passive and Active Network Measurement (PAM)*, Springer-Verlag, 2008.

[295] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, “Towards Street-level Client-independent IP Geolocation,” in *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, pp. 365–379, USENIX Association, 2011.

[296] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, “Towards IP Geolocation Using Delay and Topology Measurements,” in *Proceedings of the SIGCOMM Conference on Internet Measurement (IMC)*, pp. 71–84, ACM, 2006.

[297] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, “Constraint-based Geolocation of Internet Hosts,” *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 1219–1232, 2006.

[298] L. Stiemert and K. Gottschalk, “Geolokalisation und Verifikation von IP-Adressen,” *Bachelor Thesis*, 2012.

[299] D. Li, J. Cheny, C. Guo, Y. Liu, J. Zhangy, Z. Zhang, and Y. Zhang, “IP-Geolocation Mapping for Involving Moderately-Connected Internet Regions,” tech. rep., Microsoft Research, 2009.

[300] P. Tarrío, A. M. Bernardos, and J. R. Casar, “Weighted Least Squares Techniques for Improved Received Signal Strength Based Localization,” *Sensors*, vol. 11, no. 9, pp. 8569–8592, 2011.

[301] G. Yao, J. Bi, and A. V. Vasilakos, “Passive IP Traceback: Disclosing the Locations of IP Spoofers From Path Backscatter,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 471–484, 2015.

[302] Center for Applied Internet Data Analysis, “gig-br ark monitor statistics,” *University of California’s San Diego Supercomputer Center*, o. J. `http:www.caida.org/projects/ark/statistics/gig-br.html` (2017-02-14).

[303] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, and C. Diot, “Measurement and Analysis of Single-hop Delay on an IP Backbone Network,” *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, pp. 908–921, 2006.

[304] C. J. Bovy, H. T. Mertodimedjo, G. Hooghiemstra, H. Uijtervaal, and P. V. Mieghem, “Analysis of End-to-end Delay Measurements in internet,” in *Proceedings of the Passive and Active Measurement Workshop (PAM)*, 2002.

[305] Kowoma.de, “Geodäsie & Karten: Großkreise, Orthodrom, Loxodrom,” 2005. `http://www.kowoma.de/gps/geo/grosskreise.htm` (2017-02-14).

[306] H. Beyer, “The danger of geographic coordinates in geospatial analyses.” Australian Research Council DECRA, o. J. `http://www.spatialecology.com/beyer/blog/2013-12-03-geographic-coordinates-warning.html` (2017-02-14).

[307] J. Rumbaugh, I. Jacobson, and G. Booch, *The Unified Modeling Language Reference Manual*. Addison-Wesley Longman, 1998.

[308] E. Gamma, R. Helm, R. E. Johnson, and J. Vlissides, *Design Patterns. Elements of Reusable Object-Oriented Software*, vol. 1. Prentice Hall, 1994.

[309] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: An Open Source Software for Exploring and Manipulating Networks,” in *International AAAI Conference on Weblogs and Social Media*, 2009.

[310] E. W. Zegura, K. L. Calvert, and S. Bhattacharjee, “How to Model an Internetwork,” in *Proceedings of the Fifteenth Annual Joint Conference of the Computer and Communications Societies (INFOCOM)*, vol. 2, pp. 594–602, IEEE Computer Society, 1996.

[311] J. Winick and S. Jamin, “Inet-3.0: Internet topology generator,” tech. rep., University Michigan, 2002.

[312] A. Medina, I. Matta, and J. Byers, “BRITE: A Flexible Generator of Internet Topologies,” tech. rep., Boston University, MA, USA, 2000.

[313] M. B. Doar and A. Nexion, “A Better Model for Generating Test Networks,” in *Proceedings of Globecom*, pp. 86–93, 1996.

[314] E. Schubert, “Example data sets for ELKI,” 2015. `http://elki.dbs.ifi.lmu.de/wiki/DataSets` (2017-02-14).

[315] Czech Statistical Office, “Tschechische Republik: Regionen und größere Städte.” *Städte mit mehr als 20.000 Einwohnern*, 2015.

[316] City Population, “Niederlande: Provinzen und größere Siedlungen.” *Städte und Gemeinden mit mehr als 50.000 Einwohnern*, 2008.

[317] Bundesamt für Statistik, “Gemeindeverzeichnis (Jahresausgabe).” *Städte mit mehr als 100.000 Einwohnern*, 2015.

[318] City Population, “Europe: Population Statistics in Maps and Charts for all Countries in Europe.” *Städte Europas mit mehr als 500.000 Einwohnern*, 2015.

[319] City Population, “Frankreich: Regionen und Großstädte.” *Städte mit mehr als 50.000 Einwohnern*, 1999.

[320] Bundesamt für Statistik, “Bilanz der ständigen Wohnbevölkerung nach Bezirken und Gemeinden.” *Städte mit mindestens 10.000 Einwohnern*, 2015.

[321] Istituto Nazionale di Statistica (ISTAT), “Popolazione.” *Städte mit mehr als 50.000 Einwohnern*, 2016.

[322] City Population, “Großbritannien: Landesteile und größere Städte.” *Städtische Siedlungen mit mehr als 50.000 Einwohnern*, 2015.

[323] Census of India, “Provisional Population Totals: Cities having population 1 Lakh and above.” *Städte mit mehr als 200.000 Einwohnern*, 2011.

[324] United States Census Bureau, “cities & towns.” *Städte mit dem Status einer City über 100.000 Einwohner*, 2010.

[325] J. Choi, S. H. Noh, S. L. Min, and Y. Cho, “Towards Application / File-Level Characterization of Block References: A Case for Fine-Grained Buffer Management,” *Proceedings of ACM SIGMETRICS Conference on Measuring and Modeling of Computer Systems*, pp. 286–295, 2000.

[326] J. Choi, S. H. Noh, S. L. Min, and Y. Cho, “An Implementation Study of a Detection-Based Adaptive Block Replacement Scheme,” *Proceedings of the Annual USENIX Technical Conference*, pp. 239–252, 1999.

[327] D. Lee, J. Choi, J.-H. Kim, S. H. Noh, S. L. Min, Y. Cho, and C. S. Kim, "On the Existence of a Spectrum of Policies that Subsumes the Least Recently Used (LRU) and Least Frequently Used (LFU) Policies," *Proceedings of ACM SIGMETRICS Conference on Measuring and Modeling of Computer Systems*, pp. 134–143, 1999.

[328] T. Uhlig, P. Hillmann, and O. Rose, "Evaluation of the general applicability of Dragoon for the k-center Problem," in *Winter Simulation Conference (WSC)*, pp. 3186–3187, 2016.

[329] Réseaux IP Européens Network Coordination Centre (RIPE NCC), "RIPE Atlas," o. J. `https:atlas.ripe.net/` (2017-02-14).

[330] Trilateration Utility for Locating IP Hosts (TULIP), "What is TULIP," o. J. `http:tulip.slac.stanford.edu/` (2017-02-14).

[331] S. Laki, P. Matray, P. Haga, I. Casabai, I. Csabai, and G. Vattay, "A detailed path-latency model for router geolocation," in *Proceedings of the IEEE TridentCom*, pp. 1–6, 2009.

[332] D. S. Johnson, "A Brief History of NP-Completeness," *Journal of the Deutsche Mathematiker-Vereinigung (DMV)*, pp. 359–376, 2012.

[333] D. S. Hochbaum, G. Ausiello, and M. Lucertini, *Analysis and Design of Algorithms for Combinatorial Problems: North-Holland Mathematics Studies*, vol. 109, ch. Annals of Discrete Mathematics: Easy Solutions for the K-Center Problem or the Dominating Set Problem on Random Graphs, pp. 189–210. Elsevier Science Publishers B.V (North-Holland), 1985.

[334] V. Kann, "On the Approximability of NP-complete Optimization Problems," *PhD thesis, Department of Numerical Analysis and Computing Science, Royal Institute of Technology*, pp. 1–168, 1992.

[335] R. M. Karp, "Reducibility Among Combinatorial Problems," *R. E. Miller, J. W. Thatcher (Hrsg.): Complexity of Computer Computations. Plenum Press*, pp. 85–103, 1972.

[336] S. A. Cook, "The Complexity of Theorem-proving Procedures," in *Proceedings of the Third Annual Symposium on Theory of Computing (STOC)*, pp. 151–158, ACM, 1971.

[337] L. Levin, "Universal sorting problems," *Problems of Information Transmission, Jg. 9*, pp. 265–266, 1973.

[338] P. Sanders and R. van Stee, "Approximations- und Online-Algorithmen: k-center Problem," 2007.

[339] S. Khuller, "CMSC 651 Advanced Algorithms: 26 NP-Completeness," 1995.

[340] W.-L. Hsu and G. L. Nemhauser, *Discrete Applied Mathematics*, vol. 1, ch. Easy and hard bottleneck location problems, pp. 209–215. Elsevier Science Publishers B.V (North-Holland), 3 ed., 1979.

[341] D. S. Hochbaum, *Annals of Operations Research*, vol. 1 of *3*, ch. When Are NP-Hard Location Problems easy?, pp. 201–214. Springer US, 1984.

[342] D. S. Hochbaum and D. B. Shmoys, “A Unified Approach to Approximation Algorithms for Bottleneck Problems,” *Journal of the ACM*, vol. 33, no. 3, pp. 533–550, 1986.

[343] S. Sahni and T. Gonzalez, “P-Complete Approximation Problems,” *Journal of the ACM*, vol. 23, no. 3, pp. 555–565, 1976.

[344] I. L. Gørtz and A. Wirth, “Asymmetry in k-Center Variants,” *IT University of Copenhagen*, 2003.

Abbildungsverzeichnis

1.1 Nutzung einer stochastischen Methode für die Standortbestimmung durch Bischof Absalon vor über 850 Jahren zur Gründung der Stadt Kopenhagen 1
1.2 Transportleistung in Deutschland nach Verkehrszweigen 3
1.3 Prognostizierter Anstieg des globalen Datenverkehrs von Privatpersonen . 3
1.4 Einordnung dieser Arbeit im Zusammenhang mit weiteren Planungsbereichen bei Transportketten 7
1.5 Erweiterung von einer einstufigen zu einer zweistufigen Transportkette im Bereich Logistik 9
1.6 Erweiterung von einer einstufigen zu einer zweistufigen Transportkette im Bereich Rechnernetze 10
1.7 Szenario eines FLP anhand der 143 größten Städte in der Schweiz mit der Planung von 10 zentralen Zwischenlagern 12
1.8 Beispiel eines CDN anhand des Rechnernetzes in der Tschechischen Republik mit 124 Knoten und der Selektion von 10 Standorten zur Anbindung der Server 13
1.9 Generalisierung der spezifischen Transportketten hin zu einer generischen Modellierung unabhängig vom Anwendungsgebiet 15
1.10 Aufbau und Vorgehensweise bei der Beschreibung der vorliegenden Arbeit 22

2.1 Problem der Entwicklung eines generischen Modells für mehrere Anwendungsgebiete 23
2.2 Wechselseitiger Einfluss der vier Optimierungsaspekte 28
2.3 Gegenüberstellung einer optimierten Lösung zu dem k-center und k-median Problem mit der Platzierung von fünf Zentren 29
2.4 Beispiel des k-cluster Problems 31

3.1 Struktur des generischen Modellierungsansatzes zur Optimierung von Problemstellungen aus den Bereichen FLP und CDN sowie weiterer Anwendungsgebiete von zweistufigen Transportketten 35

4.1 Bewertung des Stands der Technik anhand von Erfolgsfaktoren im Bereich von FLP 52
4.2 Bewertung von Modellen anhand von Kriterien zur Kooperation im Bereich von CDN 52
4.3 Die ersten drei Iterationen des Zwei-Approx anhand des Beispielszenarios Deutschland aus der Logistik 55

5.1 Vorgehensweise bei der Beschreibung des Modellierungsansatzes 63
5.2 Programmstrukturplan der verschiedenen Teilprozesse des Systems für die Berechnung einer optimierten Lösung 64

5.3 Bestandteile des abstrahierten Szenarios einer Transportkette gemäß UML Klassendiagramm . . . 66
5.4 Die ersten 5000 aufgezeichneten Einträge des Traces *sprite* von über 130.000 . . . 71
5.5 Verteilung von Wörtern in Goethes Werk Faust I im Vergleich zur perfekten Zipf-Verteilung mit $\alpha = 1$ . . . 71
5.6 Vergleich von realen Daten und der Zipf-Verteilung bei referenzierten Webseiten entsprechend deren Datenverkehr und der Anzahl der Besucher 72
5.7 Generisches Modell des abstrahierten Szenarios einer zweistufigen Transportkette zur Planung und Optimierung von Zentren . . . 76
5.8 Auffassung der verschiedenen Nebenbedingungen als Schichten, welche sich additiv auf die Platzierung auswirken. . . . 78
5.9 Negative Korrelation . . . 80
5.10 Positive Korrelation . . . 80
5.11 Beispiel für die Problemstellung einer optimierten Anfragezuweisung . . . 83
5.12 Anwendung des generischen Modells auf ein spezifisches Szenarios der Logistik mit der Visualisierung der Ergebnisse in dem Simulationswerkzeug Anylogic . . . 95
5.13 Backbone Infrastruktur der Dial Telecom in der Tschechischen Republik mit Anbindung zu weiteren Ländern . . . 98
5.14 Problem des Vergleichs verschiedener Netzverbindungen von einem Nutzer zu einem Server . . . 99
5.15 Beispiel für eine IP Geolokalisation über aktive Messungen . . . 101
5.16 Prozessschritte des Modells zur Geolokalisation . . . 104
5.17 Backbone Netz von Europa . . . 106
5.18 Angenäherte Abbildungsfunktion von Zeit und Entfernung mit der Summe der kleinsten Quadrate . . . 108
5.19 Gegenüberstellung der Entfernungsbrechnung von Frankfurt am Main nach Los Angeles gemäß einer direkten Verbindung und der orthodromen Linie . 109
5.20 Visualisierung der Standortbestimmung über Lateration durch mehrere VPs 110
5.21 Unterschied zwischen Longitude und Latitude . . . 111
5.22 Verhältnis zwischen km und Grad entlang der Latitude . . . 111
5.23 Kein Schnittpunkt bei nicht überlappenden Kreisen . . . 111
5.24 Kein Schnittpunkt bei überlagerten Kreisen . . . 112
5.25 Ein Schnittpunkt bei angrenzenden Kreisen . . . 112
5.26 Zwei Schnittpunkte bei sich schneidenden Kreisen . . . 112
5.27 Lateration mit mehreren Kreisen und Schnittpunkten . . . 112
5.28 Punktwolke der Schnittpunkte mit einer Region dichter Standorte . . . 114
5.29 Automatische Anpassung der Kreisradien zum Ausgleich von Messungenauigkeiten . . . 114
5.30 TwoApprox mit der Platzierung von zehn Zentren . . . 117
5.31 MacQueen mit der Platzierung von zehn Zentren . . . 117
5.32 Lloyd mit der Platzierung von zehn Zentren . . . 117
5.33 Greedy mit der Platzierung von zehn Zentren . . . 118
5.34 Dragoon mit der Initialisierung von zehn Zentren . . . 119

5.35 Erste Iteration der Optimierung von Dragoon mit der Nebenbedingung knotenbasierte Platzierung . . . 122
5.36 Optimierung mit Dragoon mit der Nebenbedingung knotenbasierte Platzierung . . . 123
5.37 Verschiebung des aktuellen Standortes zu den Möglichkeiten der direkt benachbarten Knoten . . . 124
5.38 Erste Iteration der Optimierung von Dragoon mit der Nebenbedingung infrastrukturbasierte Platzierung . . . 125
5.39 Optimierung mit Dragoon mit der Nebenbedingung infrastrukturbasierte Platzierung . . . 125
5.40 Iterative Optimierung des Algorithmus Dragoon mit der Nebenbedingung freie Platzierung . . . 127
5.41 Erste Iteration der Optimierung von Dragoon mit der Nebenbedingung freie Platzierung . . . 127
5.42 Optimierung mit Dragoon mit der Nebenbedingung freie Platzierung . . . 128

6.1 Abstraktes UML-Klassendiagramm von den wesentlichen Bestandteilen des generischen Modells . . . 132
6.2 Abstraktes UML-Klassendiagramm von den verschiedenen Fabrikmethoden zur Erzeugung der Algorithmen, Strategien und Funktionen mittels Generator . . . 132
6.3 Abstraktes UML-Klassendiagramm von den wesentlichen Bestandteilen der Lösungsberechnung zu einem Optimierungsproblem . . . 133
6.4 Abstraktes UML-Klassendiagramm vom Zentrum mit sukzessive erweiterten Informationen über die verschiedenen Bearbeitungsschritte hinweg . . . 133
6.5 Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand der Testszenarien Logistik mit der Entfernungsberechnung gemäß Google Maps . . . 139
6.6 Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand der Testszenarien Rechnernetze mit der Entfernungsberechnung über Dijkstra auf dem Graphen . . . 139
6.7 Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand der Testszenarien Synthetisch mit der Entfernungsberechnung gemäß Euklid . . . 140
6.8 Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand des Szenarios Dialtelecom mit der Entfernungsberechnung gemäß Google Maps . . . 141
6.9 Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand des Szenarios Dialtelecom mit der Entfernungsberechnung über Dijkstra auf dem Graphen . . . 141
6.10 Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand des Szenarios Dialtelecom mit der Entfernungsberechnung gemäß Euklid . . . 142
6.11 Vergleich der Platzierungsleistung bei verschiedener Anzahl von Zentren anhand des Szenarios Dialtelecom mit der Entfernungsberechnung gemäß Orthodrome . . . 142

6.12 Abweichung der Leistung vom theoretischen Optimum hinsichtlich des Kriteriums der maximalen Entfernung . . . 143
6.13 Einfluss von Backtracking auf die Platzierungsleistung ungeachtet des steigenden Berechnungsaufwandes . . . 144
6.14 Vergleich der Optimierungsleistung bei der Platzierung von Zentren unter Beachtung der Nebenbedingung knotenbasierte und freie Platzierung . . . 148
6.15 Einfluss unterschiedlicher Zielkriterien bei der Optimierung auf die maximale Entfernung bei den Nebenbedingungen knotenbasierte und freie Platzierung . . . 149
6.16 Einfluss unterschiedlicher Zielkriterien bei der Optimierung auf die durchschnittliche Entfernung bei den Nebenbedingungen knotenbasierte und freie Platzierung . . . 150
6.17 Durchschnittlicher Berechnungsaufwand für eine Ausführung des jeweiligen Algorithmus von den Referenzalgorithmen . . . 151
6.18 Durchschnittlicher Berechnungsaufwand für eine Ausführung des jeweiligen Algorithmus gegenüber den verschiedenen Varianten von Dragoon . . . 152
6.19 Durchschnittlicher Berechnungsaufwand für eine Ausführung des jeweiligen Algorithmus bei unterschiedlicher Anordnung und Anzahl von Nutzern für 10 Zentren . . . 153
6.20 Durchschnittlicher Berechnungsaufwand für eine Ausführung des jeweiligen Algorithmus bei unterschiedlicher Anordnung und Anzahl von Nutzern für 30 Zentren . . . 153
6.21 Leistungsübersicht hinsichtlich der maximalen Entfernung: Die orange, vertikale Linie markiert die Anzahl von 25 Zentren, wobei die durchschnittliche Verbesserung danach weniger als 2 % pro zusätzlichem Zentrum beträgt. . . 156
6.22 Kostenoptimierung am Szenario Schweiz . . . 157
6.23 Leistungsvergleich von Speicherersetzungsstrategien gegenüber dem Optimum mittels Bélády bei verschiedenen Speichergrößen und 5 % Anteil der Nutzerprofile vom Universum . . . 162
6.24 Leistungsvergleich von Speicherersetzungsstrategien gegenüber dem Optimum mittels Bélády bei verschiedenen Speichergrößen und 15 % Anteil der Nutzerprofile vom Universum . . . 163
6.25 Einfluss der Speicherkapazität auf die Fehlerrate bei der Speicherverwaltung mittels Bélády . . . 164
6.26 Einfluss der Speicherkapazität auf die Fehlerrate bei der Speicherverwaltung mittels LRU-2 . . . 165
6.27 Zusammenhang von Korrelation der Nutzerprofile und der Fehlerrate bei Anfragen . . . 166
6.28 Pareto-Front bei der Optimierung von der Entfernung und der Profilkorrelation mit 5 Zentren . . . 168
6.29 Pareto-Front bei der Optimierung von der Entfernung und der Profilkorrelation mit 10 Zentren . . . 168

A.1 Gegenüberstellung von Extremfällen zur Motivation für den Einsatz von Zentren . . . 215

B.1 Beweiskette für die NP-schwere des k-center und k-cluster Problems . . . 219

C.1 Kein Zentrum $z^* \in Z$ einer optimalen Lösung von Z^* darf in der Hülle von zwei unterschiedlichen Zentren sein. 224

D.1 Zufall Rechteck 1, 100 Knoten 225
D.2 Zufall Rechteck 2, 100 Knoten 225
D.3 Zufall Rechteck 3, 200 Knoten 226
D.4 Zufall Quadrat, 200 Knoten 226
D.5 Gephi 1, 100 Knoten 226
D.6 Gephi 2, 200 Knoten 226
D.7 Gephi 3, 200 Knoten 227
D.8 Gephi 4, 300 Knoten 227
D.9 Elki density, 150 Knoten 227
D.10 Elki Mouse, 500 Knoten 227
D.11 Tschechien, 62 Knoten 228
D.12 Niederlande, 67 Knoten 228
D.13 Deutschland, 77 Knoten 228
D.14 Europa, 99 Knoten 228
D.15 Frankreich, 114 Knoten 229
D.16 Schweiz, 144 Knoten 229
D.17 Italien, 147 Knoten 229
D.18 Vereinigtes Königreich, 196 Knoten 229
D.19 Indien, 230 Knoten 230
D.20 United States of America, 273 Knoten 230
D.21 Latnet, 68 Knoten und 73 Kanten 230
D.22 Uninett, 47 Knoten und 102 Kanten 230
D.23 DialtelecomCz, 124 Knoten und 126 Kanten 231
D.24 GtsCe, 141 Knoten und 181 Kanten 231
D.25 TataNld, 143 Knoten und 181 Kanten 231
D.26 Colt, 146 Knoten und 164 Kanten 231
D.27 US Carrier, 151 Knoten und 174 Kanten 232
D.28 Cogentco, 180 Knoten und 210 Kanten 232
D.29 Westeuropa, 271 Knoten und 385 Kanten 232
D.30 Europa, 447 Knoten und 605 Kanten 232

E.1 Vergleich der Leistung von Llyod bei der Platzierung von Zentren unter Beachtung der Nebenbedingung knotenbasierte und freie Platzierung . . . 233
E.2 Vergleich der Leistung von MacQueen bei der Platzierung von Zentren unter Beachtung der Nebenbedingung knotenbasierte und freie Platzierung 234

F.1 Abbildungsfunktion gemäß der Korrelation von Latenz und Entfernungen nach Google Maps 238
F.2 Abbildungsfunktion gemäß der Korrelation von Latenz und orthodromen Entfernungen 238

G.1 Übersicht der Vorabveröffentlichungen und Einordnung in das Gesamtkonzept der Arbeit 243

Tabellenverzeichnis

1.1 Gegenüberstellung der verschiedenen Begriffe bei Transportketten den Domänen Logistik und Rechnernetze mit entsprechender generischer Bezeichnung . . . 17

4.1 Parameter der Simulation und der Zipf-Verteilung . . . 47
4.2 Übersicht bisheriger Arbeiten bewertet entsprechend der Vergleichseigenschaften . . . 51
4.3 Beispiel-Matrix des Greedy Algorithmus mit 5 Knoten . . . 56

5.1 Beispiel einer Anfrageverteilung von einem Nutzer . . . 69
5.2 Beispiel eines Traces anhand der ersten, aufgezeichneten Referenzen des Datensatzes *multi2* . . . 70
5.3 Bewertung verschiedener Standorte anhand mehrerer Nebenbedingungen 79
5.4 Beispiel zur Rangkorrelation gemäß Pearson . . . 81

6.1 Testszenarien der Kategorie Synthetisch . . . 134
6.2 Testszenarien der Kategorie Logistik gemäß FLP . . . 135
6.3 Testszenarien der Kategorie Rechnernetz gemäß CDN . . . 135
6.4 Testdaten der Experimente . . . 146
6.5 Durchschnittliche Leistung bei der knotenbasierten Platzierung von alternativen Algorithmen im Vergleich zu Dragoon (ΔD_{avg}) basierend auf 2000 zufälligen Szenarien . . . 146
6.6 Gegenüberstellung der Δ D Leistung mehrerer Algorithmen am Beispiel von Problemstellungen mit 49 Nutzern und 4 Zentren . . . 147
6.7 Statistische Auswertung der Leistungsdifferenzen von der Platzierung bei den Nebenbedingungen Knotenplatzierung und freie Platzierung . . . 148
6.8 Leistungsvergleich für das Zielkriterium Minimierung der maximalen Entfernung in % . . . 155
6.9 Verbesserung der Entfernung in Relation zur Anzahl der Zentren mittels Dragoon anhand der Knotenplatzierung . . . 155
6.10 Vergleich der Genauigkeit des spezifischen Modells zur IP Geolokalisation mit 20 VPs bei Verwendung unterschiedlicher Landmark Positionen und Approximationen für die Entfernung . . . 159
6.11 Vergleich der Genauigkeit des spezifischen Modells zur IP Geolokalisation mit 20 VPs gegenüber anderen aktiven und passiven Diensten . . . 160
6.12 Parameter des Experiments Speicherersetzungsstrategien . . . 161
6.13 Parameter des Experiments Speicherverhalten . . . 163
6.14 Parameter des Experiments Leistungspotential . . . 166

6.15 Gegenüberstellung der Leistungswerte bei den Zielen kurzer Entfernung und hoher Korrelation bei 20 Zentren 166
6.16 Parameter des Experiments Pareto-Front 169

A.1 Gegenüberstellung ausgewählter Fälle der Platzierung von Zentren und resultierenden Auswirkungen 217

E.1 Statistische Auswertung der Leistungsdifferenzen von Llyod bei der Platzierung bei den Nebenbedingungen Knotenplatzierung und freie Platzierung 234
E.2 Statistische Auswertung der Leistungsdifferenzen von MacQueen bei der Platzierung bei den Nebenbedingungen Knotenplatzierung und freie Platzierung 234
E.3 Leistungsvergleich für das Zielkriterium Minimierung der maximalen Entfernung in % 235

F.1 Datensatz der verwendeten Ziele zur Überprüfung der IP Geolokalisation . 239

Abkürzungs- und Symbolverzeichnis

α ..... Zipf Parameter 47
β ..... Approximationsgüte 54
δ ..... Anzahl von Backtracking Schritten beim Greedy Algorithmus 56
Δ D ..... Leistungsdifferenz 146
γ ..... Problemspezifischer Faktor von GRASP 57
λ ..... Faktor der Abbildungsfunktion im Modell der Geolokalisation 107
$\mathrm{Cov}(X, Y)$ ... Kovarianz 80
$\overline{rang}_x$ ..... Mittelwert der Ränge von x 81
ρ ..... Korrelationsfunktion 31
$\rho_{X,Y}$ ..... Korrelatioskoeffizient nach Pearson 80
σ_{rang_x} ..... Standardabweichung der Ränge von x 81
τ ..... Korrelatioskoeffizient nach Kendall 80
$d(v_i)$ ..... Zielfunktionswert von einem Knoten des Graphen 30
r_s ..... Korrelatioskoeffizient nach Spearman 80
ARC ..... engl. Adaptive Replacement Cache 89
ASN ..... engl. Autonomous System Number 239
AVG ..... engl. Average; Durchschnitt 147
C ..... Menge aller Korrelationskoeffizienten eines Nutzers 31
CAR ..... engl. Clock with Adaptive Replacement 89
CBG ..... engl. Constraint-Based Geolocation 101
CCN ..... engl. Content Centric Networking 41
CDN ..... engl. Content Delivery Network 10
D ..... Menge aller Zielfunktionswerte des Graphen 30
d ..... Zielfunktion 30
d(r) ..... Güte der Lösung einer Optimierungsaufgabe 53
d* ..... Güte der optimalen Lösung einer Optimierungsaufgabe 53
DNS ..... engl. Domain Name System 100
E ..... Menge von Verbindungen eines Graphen 30
e ..... Kante 30
ELKI ..... engl. Environment for Developing KDD-Applikations Supported by Index-Structures 61
EW ..... Erwartungswert 81
FBR ..... engl. Frequency-Based Replacement 89

FIFO ..... engl. First In First Out 89
FLP ..... engl. Facility Location Problem 9
FQDN ..... engl. Fully Qualified Domain Name 100
FTP ..... engl. File Transfer Protocol 42
G ..... Graph 30
GM ..... Google Maps 158
GRASP ..... engl. Greedy Randomized Adaptive Search Procedures 56
HaS ..... engl. Hub and Spoke Systems, Nabe-Speiche-Systeme 40
HYP ..... Hyperbolisches Verfahren mit minimaler quadratischer Abschätzung . 158
IANA ..... engl. Internet Assigned Numbers Authority 100
ICMP ..... engl. Internet Control Message Protocol 106
ICN ..... engl. Information-Centric Networking 48
ID ..... Identifikator 67
ILP ..... engl. Integer Linear Programming 59
IP ..... Internet Protokoll 99
ISO ..... engl. International Organization for Standardization 16
ISP ..... engl. Internet-Service-Provider 12
ITU ..... engl. International Telecommunication Union 17
K ..... Menge der Zentren 30
L(p) ..... Lösungsraum einer Optimierungsaufgabe 53
LAC ..... engl. Latency-Aware Caching 41
LFU ..... engl. Least Frequently Used 89
LIRS ..... engl. Low Inter-reference Recency Set 90
LP ..... engl. Linear Programming 59
LRFU ..... engl. Least Recently/Frequently Used 89
LRU ..... engl. Least Recently Used 89
MAX ..... Maximum 147
MOCell ..... engl. Multiobjective Cellular Genetic Algorithm 62
MOEA ..... engl. Multiobjective Evolutionary Algorithms 62
MQ ..... engl. Multi Queue 89
MRU ..... engl. Most Recently Used 90
n ..... Länge der Verkoren X, Y 81
NSGAII ..... engl. Non-dominated Sorting Genetic Algorithm 62
OD ..... Orthodrome Distanz 158
p ..... Optimierungsaufgabe 53
PAES ..... engl. Pareto Archived Evolution Strategy 62
pf ..... Interessen-Profile 31
QoE ..... engl. Quality-of-Experience; Erfahrungserlebnis 13
QoS ..... engl. Quality-of-Service; Dienstqualität 13

Qx ..... x-te Quantil 79
r ..... Lösung einer Optimierungsaufgabe 53
rang(x) ..... Rang von x 81
RN ..... Rechnernetze 50
RTT ..... engl. Round-Trip-Time 103
SC ..... engl. Supply Chain 50
SDN ..... engl. Software Defined Networking 15
SPEA2 ..... engl. Strength-based Evolutionary Algorithm 62
TBG ..... engl. Topology-Based Geolocation 101
TCP ..... engl. Transmission Control Protocol 106
UDP ..... engl. User Datagram Protocol 106
V ..... Menge von Knoten eines Graphen 30
v ..... Knoten 30
Var ..... Varianz 81
VEGA ..... engl. Vector Evaluated Genetic Algorithm 62
VP ..... engl. Vantage Point 100
WEKA ..... engl. Waikato Environment for Knowledge Analysis 61
X ..... 1. Vektor der Korrelation 81
Y ..... 2. Vektor der Korrelation 81

A Beispiel zur Motivation einer strategischen Planung von Zentren

Zur Veranschaulichung der Problemstellung werden in diesem Abschnitt an einem Beispiel drei Fälle unterschiedlicher Platzierung von Zentren analysiert. In dem betrachteten Beispiel gibt es drei Nutzer und fünf Quellen. Die Teilnehmer des Systems sind wie in Abbildung A.1 miteinander verbunden. Die Quellen in der Domäne bieten die verschiedenen Dienste A, B, C, D und E an, wobei alle Dienste vergleichbare Eigenschaften besitzen und sich vervielfältigen lassen. Die Nutzer fragen die Dienste in deren Interessen-Profil zu gleicher Intensität, Häufigkeit und Wahrscheinlichkeit an. Alle Nutzer haben Interesse am Dienst A und B, während Nutzer 1 zusätzlich Dienst C, Nutzer 2 Dienst D und Nutzer 3 Dienst E im Profil aufweist. Zur Veranschaulichung des Problems stellen die Nutzer zu einem Zeitpunkt gleichzeitig deren Anfragen nach Diensten an das System. Diese treffen in zufälliger Reihenfolge bei dem jeweiligen Zentrum oder der jeweiligen Quelle ein. Ein Zentrum ist auf Grund der begrenzten Ressourcen nicht in der Lage sämtliche Dienste der Domäne zwischenzuspeichern und muss somit seinen verfügbaren Speicher entsprechend verwalten. Die Zentren besitzen in dem Beispiel eine Speicherkapazität von zwei Diensten und müssen zur Beantwortung der Anfragen den jeweiligen Dienst zuvor im Speicher haben. Dafür fordert das jeweilige Zentrum bei Bedarf den Dienst bei der Quelle an.

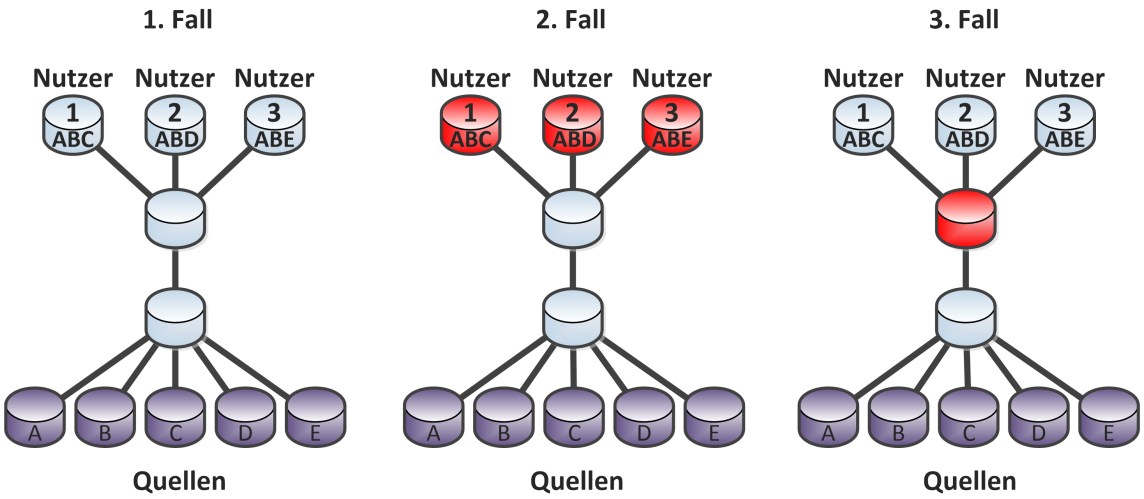


Abbildung A.1: Gegenüberstellung von Extremfällen zur Motivation für den Einsatz von Zentren

Die drei betrachteten Fälle unterscheiden sich in der Anzahl und Position der Zentren. Die Fälle sind:

1. Kein Zentrum: Alle Anfragen der Nutzer sind an die Quelle zu stellen.
2. Zentrum bei jedem Nutzer: Jeder Nutzer fragt sein persönliches Zentrum nach einem Dienst an.
3. Zwischenlösung: Strategisch platzierte Zentren werden durch die Nutzer angefragt.

1. Fall - Kein Zentrum

Wenn kein Zentrum zur Zwischenspeicherung der Dienste existiert, dann müssen die Nutzer stets die Quellen anfragen. Dadurch wird eine Anfrage stets positiv beantwortet. In diesem Beispiel haben die Nutzer die längst möglichen Verbindungsstrecken von jeweils 3 Verbindungslängen zu den Quellen. Durch die langen Verbindungen entsteht eine hohe Transferzeit bei Anfragen und ein möglicher Flaschenhals bei der mittleren Verbindung. Somit wird die Infrastruktur stark belastet und die Nutzer erhalten die Dienste mit hoher Verzögerung. Eine Aktualisierung der Dienste wirkt sich nicht auf das Verhalten des Gesamtsystems aus.

2. Fall - Zentrum bei jedem Nutzer

Ungeachtet der notwendigen Menge an Ressourcen erhält in diesem Fall jeder Nutzer zur Leistungssteigerung sein eigenes Zentrum. Die Nutzer stellen hierbei die Anfragen stets an das jeweilige nächstgelegene Zentrum. Durch den Einsatz der drei Zentren reduziert sich die Verbindungsstrecke für eine Anfrage des Nutzers somit auf das theoretische Minimum von 0 Verbindungslängen. Da ein Zentrum nur 2 Dienste gleichzeitig zwischenspeichern kann, haben alle Nutzer bei optimaler Verwaltung des Speichers durchschnittlich eine $66,\overline{6}$ % Wahrscheinlichkeit, dass ein Zentrum die Anfrage sofort positiv beantwortet. Anderenfalls ist eine Verbindungsstrecke von der Länge 3 erforderlich. Somit ergibt sich für das Gesamtsystem eine durchschnittliche Länge von 1 für eine Anfrage eines Nutzers. Sobald sich die Dienste ändern, müssen diese erneut von den Zentren bei den Quellen mit der Länge von 3 Verbindungen angefordert werden. Dies erhöht die durchschnittliche Anfragelänge.

3. Fall - Zwischenlösung

Bei dieser strategischen Standortwahl des Zentrums haben alle Nutzer eine Verbindungslänge von 1 zum Zentrum. Wie im vorherigen 2. Fall, haben alle Nutzer durchschnittlich eine $66,\overline{6}$ % Wahrscheinlichkeit einer erfolgreichen Anfrage bei optimaler Speicherverwaltung, da das Zentrum nur 2 Speicherplätze (3. Fall a) besitzt. Wenn das Zentrum den Dienst nicht zur Verfügung hat, dann fordert es diesen über

2 Verbindungslängen an. Somit ergibt sich für das Gesamtsystem eine durchschnittliche Länge von $1,\overline{6}$ Verbindungslängen. Bei einer Aktualisierung der Dienste sind im Gegensatz zum 2. Fall nur 2 Verbindungslängen erforderlich. Somit fällt die Last auf die gesamte Infrastruktur mit steigender Aktualisierungshäufigkeit geringer aus.

Wenn das eine Zentrum eine Speicherkapazität von 5 hat (3. Fall b), anstatt von 2 (3. Fall a), dann ist das Zentrum in der Lage alle Dienste zwischenzuspeichern. Die durchschnittliche Verbindungslänge reduziert sich damit auf 1.

Zusammenfassung

Die Tabelle A.1 stellt die verschiedenen Fälle anhand mehrere Kriterien gegenüber. Anhand des kleinen Beispiels wird der Vorteil einer strategischen Platzierung eines Zentrums ersichtlich. Durch den gezielten Ressourceneinsatz von nur einem Zentrum mit angepasster Speicherkapazität im 3. Fall b wird für die Nutzer in der Gesamtheit die beste Situation erreicht.

Tabelle A.1: Gegenüberstellung ausgewählter Fälle der Platzierung von Zentren und resultierenden Auswirkungen

Eigenschaft	1. Fall	2. Fall	3. Fall a	3. Fall b
Anzahl der Zentren	0	3	1	1
Durchschnittliche Verbindungslänge	3	1	$1,\overline{6}$	1
Summe der Speicherkapazität der Zentren	0	6	2	5
Verbindungslänge bei Dienstaktualisierung	0	3	2	2

B Beweis der NP Vollständigkeit des Problems

Um die Schwierigkeit des *k*-center und *k*-cluster Problems zu verdeutlichen, werden diese über eine Polynomialzeitreduktion gemäß den Methoden der theoretischen Informatik auf anderen Problemen abgebildet und verglichen [332]. Durch die Reduktion wird belegt, dass jeder Algorithmus, der das reduzierte Problem löst, sich auch zur Lösungsberechnung des ursprünglichen Problems verwenden lässt. Somit haben alle Probleme einer Komplexitätsklasse eine vergleichbare Schwierigkeit hinsichtlich des Berechnungsaufwands von Algorithmen zur optimalen bzw. exakten Lösungsfindung mittels Turingmaschine.

Das *k*-center und das *k*-cluster Problem lassen sich auf die Problemstellung *Dominating Set* reduzieren [333]. Für dieses wurde die Reduktion auf das *Set Covering* Problem gezeigt [334], für welches bereits die NP-Vollständigkeit nachgewiesen wurde [335]. Dazu erfolgte wiederum eine Ableitung auf das *Erfüllbarkeitsproblem der Aussagenlogik* (engl. Satisfiability, kurz SAT), für das initial die NP-Vollständigkeit bewiesen wurde. Problemstellungen der Komplexitätsklasse NP lassen sich nur nichtdeterministisch in Polynomialzeit lösen. Das Erfüllbarkeitsproblem der Aussagenlogik ist eins der schwersten Probleme der Komplexitätsklasse NP gemäß *Satz von Cook* [336, 337]. Karp [335] hat für mehr als zwanzig Probleme der Komplexitätsklasse NP die Reduktion auf das SAT Problem gezeigt.

Die Abbildung B.1 verdeutlicht die Beweiskette zur Reduktion des *k*-center und des *k*-cluster Problems hinsichtlich der Schwierigkeit NP-vollständig.

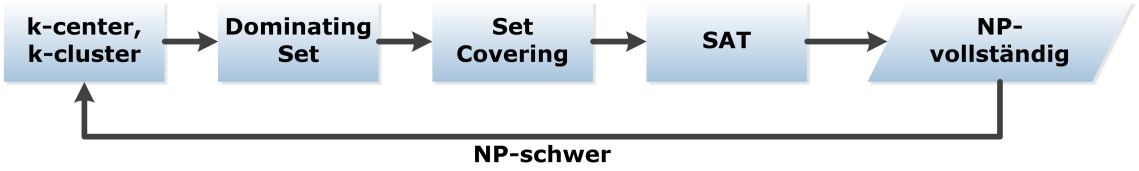


Abbildung B.1: Beweiskette für die NP-schwere des *k*-center und *k*-cluster Problems

Die Klasse NP wird weiterhin in NP-vollständig und NP-schwer unterteilt. Als NP-vollständig werden die Probleme bezeichnet, welche die schwierigsten Probleme in der Komplexitätsklasse NP sind. Alle anderen Probleme der Klasse werden mit NP-schwer bezeichnet. Diese lassen sich mit polynomialem Aufwand auf NP-vollständige Probleme ableiten. Die Lösung eines NP-schweren Problems in Polynomialzeit hätte zur Folge, dass auch jedes andere Problem in NP in polynomialer Zeit lösbar ist. Bisher konnte für kein Problem in NP nachgewiesen werden, dass es in polynomialer Zeit lösbar ist.

Der folgende Beweis erfolgt nach [65, 176, 333, 338] und beschreibt die Polynomialzeitreduktion auf das Finden einer dominierenden Menge gemäß dem von *Dominating Set* Problem.

Die dominierende Menge ist eine Teilmenge S von Knoten, sodass jeder Knoten der nicht in S ist, an einen Knoten in S angrenzt. Das Finden einer dominierenden Menge mit minimaler Größe ist NP-vollständig. [334, 339] Für einen Graph G sei dom(G) als die Größe der kleinstmöglichen, dominierenden Menge definiert. Die dominierende Menge ist vergleichbar mit der Hülle der Zentren, jedoch nicht gleichzusetzen.

Theorem:
Es ist NP-schwer das allgemeine *k*-center Problem mit einem beliebigen Faktor β zu approximieren.

Beweis:
Es erfolgt eine Reduzierung des *k*-center Problems auf das Finden einer dominierenden Menge im Graphen G = (V, E). Dazu wird ein Graph G' = (V, E = V x V) angenommen, indem jeder Knoten mit jedem verbunden ist. Weiterhin ist die Fitnessfunktion d(u,v) gegeben, wobei u und v Knoten sind:

$$d(u,v) = \begin{cases} 1 & \textit{wenn } u,\ v \in E \\ 2\beta & \textit{sonst} \end{cases} \tag{B.1}$$

Es wird angenommen, das für G eine dominierende Menge mit höchstens der Größe von *k* existiert. Dann gibt es eine Lösung mit dem Fitnesswert 1 für G'. Ein entsprechender β-approximierender Algorithmus liefert demzufolge als Ergebnis 1 mit einem Fitnesswert w $\leq \beta$ gemäß Theorem. Wenn es keine dominierende Menge in G gibt, hat jedes *k*-center Problem einen Fitnesswert w $\geq 2\beta \geq \beta$.

Angenommen es existiert ein β-approximierender Algorithmus für das *k*-center Problem, dann ist dieser Entscheidungsalgorithmus auch β-approximierend für G', da G' $\in$ G. Die Lösung für das Problem hat demzufolge ein Gewicht $w \leq \beta$. Dies bedeutet wiederum, dass eine dominierende Menge mit höchstens der Größe k existiert. Anderenfalls gibt es keine dominierende Menge. ■

Der Beweis zeigt die Reduktion des *k*-center Problems auf die Problemstellung *Dominating Set* und somit die NP-schwere. Der Beweis zeigt, dass es keinen beweisbaren Lösungsalgorithmus für das *k*-center Problem gibt, welcher eine garantierte Leistung besser als zwei-approximiert liefert. Diese gilt jedoch nur unter der Annahme, dass P $\neq$ NP gilt.

Für angelehnte Problemstellungen ist ein Beweis mit Reduktion auf das *k*-center Problem möglich. Als Beispiel dient die Problemstellung mit gewichteten Knoten und gewichteten Kanten. Dabei ist lediglich die Dreiecks-Ungleichheit nachzuweisen, d(u,w) $\leq$ d(u,v) + d(v,w) mit u,v,w $\in$ V, um auf das *k*-center Problem referenzieren zu können.

Das *k*-center Problem mit Dreiecks-Ungleichheit ist somit NP-schwer für jede β Approximierung für $\beta \leq 2$ [340, 341].

B.1 Angelehnte Problemstellungen im Überblick

Im Folgenden sind verschiedene Teilprobleme des betrachteten Anwendungsgebietes formal beschrieben. Eine entsprechende Approximierbarkeit ist jeweils mit angegeben, um das Problem zu verdeutlichen. Dies zeigt, dass die jeweiligen Teilprobleme für sich betrachtet ebenfalls in der Komplexitätsklasse NP liegen [53]. Die Lösungsberechnung erfolgt daher auf Grund der Komplexität und der Anzahl an Lösungsmöglichkeiten mittels approximierenden Ansätzen.

B.2 *k*-center

Für die in dieser Arbeit betrachteten *k*-center Probleme gilt stets die Dreiecks-Ungleichheit:

d(u,v) ist die Entfernung zwischen den Knoten u und v, wobei d(u,v) + d(v,w) $\geq$ d(u,w)

B.2.1 Minimum *k*-center ohne Gewichte

Das einfache *k*-center Problem ohne Wichtung von Knoten und Kanten lässt sich wie folgt zusammenfassen [342]:

I = {G = (V,E) sei ein vollständiger Graph, f: E $\rightarrow \mathbb{N}$ mit Kantengewichten, welche der Dreiecks-Ungleichheit genügen}
S ($< V, E, f >$) = { Z $\subseteq$ V }
m ($< V, E, f >$,Z) = $max_{v \in V}\ min_{z \in Z}$ f({ v,z })
Optimierungsziel = Minimierung von m

Approximierbarkeit:
Minimales *k*-center ist nicht approximierbar mit 2 - β für $\beta > 0$.

B.2.2 Minimum *k*-center mit Gewichten

Das *k*-center Problem mit Wichtung von Kanten lässt sich wie folgt zusammenfassen [342]:

I = {G = (V,E) sei ein vollständiger Graph, f: E $\rightarrow \mathbb{N}$ mit Kantengewichten, welche der Dreiecks-Ungleichheit genügen, w: V $\rightarrow \mathbb{N}$}
S ($< V, E, f >$,w) = { Z $\subseteq$ V }
m ($< V, E, f, w >$,Z) = $max_{v \in V}\ min_{z \in Z}$ f({ v,z })
Optimierungsziel = Minimierung von m

Approximierbarkeit:
Minimales k-center ist nicht approximierbar mit 3 - β für $\beta > 0$.

B.3 *k*-cluster

Für die Referenzierung des k-cluster Problems auf die Prolemstellung k-center muss lediglich die Dreiecks-Ungleichheit nachgewiesen werden. Bei realistischen Anwendungsfällen ist davon auszugehen.

B.3.1 Minimum *k*-clustering mit Gewichten

Die Problemstellung mit gewichteten Kanten bezieht sich auf den Anwendungsfall der Optimierung hinsichtlich der maximalen Entfernung der Nutzer zu deren jeweils nächstgelegenen Zentrum.

I = {G = (V,E) sei ein vollständiger Graph, f: E $\rightarrow \mathbb{N}$ mit Kantengewichten, welche der Dreiecks-Ungleichheit genügen}
S($< V, E, f >$) = {F = {Z_1,Z_2,...,Z_k} eine Teilmenge von V}
m($< V, E, f >$,F) = $max_{i \in [1..k]; v1, v2 \in Z_i}$ f({v_1,v_2})
Optimierungsziel = Minimierung von m

Approximierbarkeit:
Minimales k-clustering ist nicht approximierbar mit 2 - β für $\beta > 0$ [342].

B.3.2 Minimum *k*-clustering hinsichtlich Summe bzw. Durchschnitt

Die Problemstellung bezieht sich auf den Anwendungsfall der Optimierung hinsichtlich der durchschnittlichen Entfernung der Nutzer zu deren jeweils nächstgelegenen Zentrum.

I = {G = (V,E) sei ein vollständiger Graph, f: E $\rightarrow \mathbb{N}$}
S($< V, E, f >$) = {F = {Z_1,Z_2,...,Z_k} eine Teilmenge von V}
m($< V, E, f >$,F) = $\sum_{i=1}^{k} \sum_{v_1, v_2 \in Z_i}$ f({v_1,v_2})
Optimierungsziel = Minimierung von m

Approximierbarkeit:
Minimales k-clustering Sum ist nicht in der Komplexitätsklasse der Entscheidungsprobleme, welche sich von einer deterministischen Turingmaschine in exponentieller Zeit berechnen lassen. [343]

C Beweis über die garantierte Lösungsqualität

Für die *k*-center Problemstellung mit der Nebenbedingung Knotenplatzierung wurde bewiesen, dass der Greedy Algorithmus bei der Lösungsberechnung eine zwei-approximierte Lösungsgüte garantiert [65]. Dieser Beweis lässt sich entsprechend auf weitere heuristische Algorithmen übertragen, da dieser nicht spezifisch für den Greedy Ansatz ist [78, 79]. Bei der Problemstellung wird vom symmetrischen *k*-center Problem ausgegangen. Hierbei entspricht die Entfernung von einem Knoten A zu Knoten B der Entfernung von B zu A. Für die Betrachtung der asymmetrischen Varianten und einer Gegenüberstellung wird auf [344] verwiesen.

Theorem:
Eine Menge von Zentren Z, welche durch einen heuristischen Algorithmus ermittelt wurde, erreicht bei der Vorgabe von der Anzahl *k* eine garantierte Lösungsqualität von einer Hülle $R(Z) \geq 2r$.

Der folgende Beweis ist an die Arbeiten von [65, 78, 79] angelehnt:

Angenommen es gibt eine Menge von Zentren Z^* von höchstens k Zentren mit einem Radius $r(Z^*) \leq r$ entsprechend der optimalen Hülle. Jedes Zentrum $z \in Z$, welches von einem heuristischen Algorithmus selektiert wurde, gehört zu der Menge der Nutzer N und die Menge Z^* hat einen Radius von höchstens r. Dann gibt es mindestens ein Zentrum $z^* \in Z^*$, welches die höchste Entfernung r von einem Zentrum z hat. Diese ist $d(z,z^*) \leq r$. Daraus lässt sich schlussfolgern, dass ein solches Zentrum z^* nahe einem weiteren Zentrum z liegt.

Dementsprechend wird die Behauptung aufgestellt, dass kein Zentrum z^* in der optimalen Lösung von Z^* nahe eines zweiten Zentrums liegen kann. Falls dies nicht gilt, steht die Behauptung im Widerspruch zur Definition vom Radius $r(Z^*) \leq r$. Zum Nachweis der Behauptung werden die beiden Möglichkeiten überprüft.

1. Möglichkeit:
Jedes Zentrum $z \in Z$ hat ein nahe gelegenes, optimales Zentrum $z^* \in Z^*$. Die Menge dieser Zentren Z^* ist von der Menge Z verschieden. Dies hat zur Folge, dass $|Z^*| \geq |Z|$ gilt. Damit steht die Annahme $|Z| > k$ im Widerspruch zur Behauptung Z^* besitzt höchstens k Zentren.

2. Möglichkeit:
Es kann kein Zentrum $z^* \in Z$ in der optimalen Lösung von Z^* nahe zu einem weiteren Zentrum $z \in Z$ liegen. Dies ist in der Abbildung C.1 dargestellt. Demzufolge gibt es kein optimales Zentrum $z^* \in Z$ in der Nähe von zwei Zentren z, z' $\in$ Z. Jedes Paar von

Zentren z, z' $\in$ Z muss demzufolge eine Entfernung von mehr als 2r haben. Wenn ein $z^* \in$ Z innerhalb der Entfernung von höchstens r liegt, dann steht dies im Widerspruch zur Dreiecks-Ungleichheit d(z,z^*) + d(z^*,z') $\geq$ d(z,z') $>$ 2r. ■

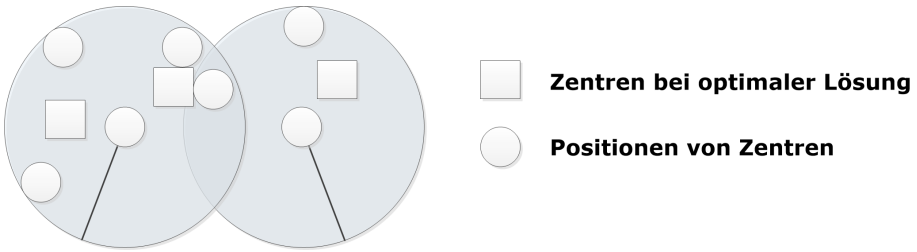


Abbildung C.1: Kein Zentrum $z^* \in$ Z einer optimalen Lösung von Z^* darf in der Hülle von zwei unterschiedlichen Zentren sein. (Vgl. [65])

D Szenarien zur Evaluation

Im Folgenden sind die verwendeten Szenarien zur Evaluation des generischen Modells abgebildet. Diese unterteilen sich in die Kategorien Synthetisch, Logistik und Rechnernetze.

D.1 Synthetisch

Die Abbildungen D.1 bis D.10 stellen die synthetischen Szenarien dar.

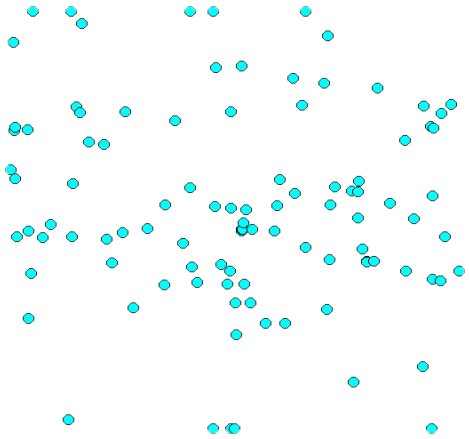


Abbildung D.1: Zufall Rechteck 1, 100 Knoten

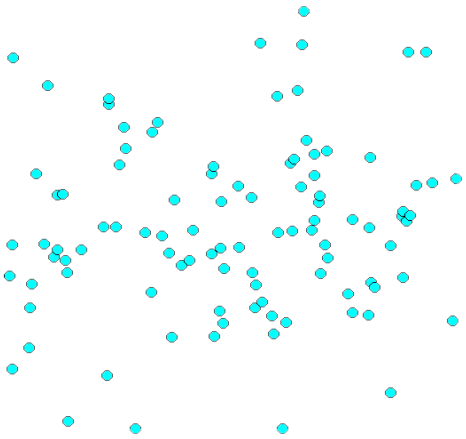


Abbildung D.2: Zufall Rechteck 2, 100 Knoten

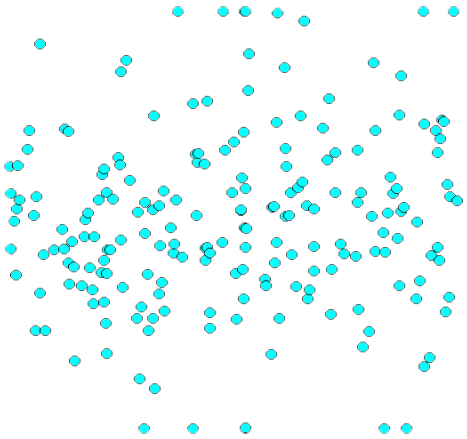


Abbildung D.3: Zufall Rechteck 3, 200 Knoten

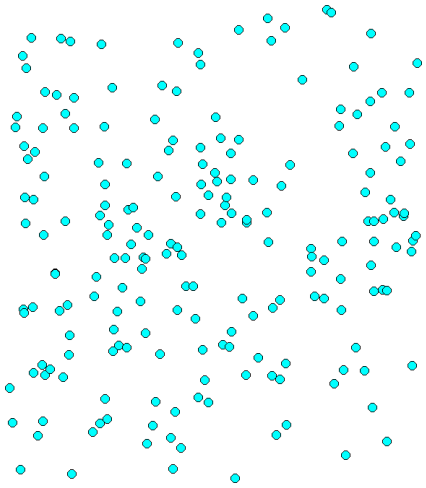


Abbildung D.4: Zufall Quadrat, 200 Knoten

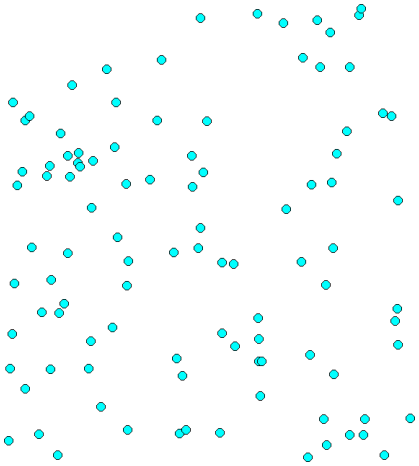


Abbildung D.5: Gephi 1, 100 Knoten

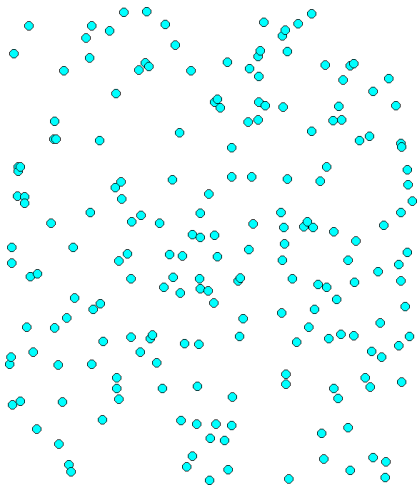


Abbildung D.6: Gephi 2, 200 Knoten

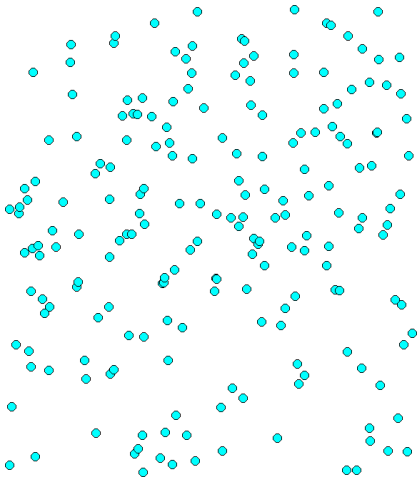


Abbildung D.7: Gephi 3,
200 Knoten

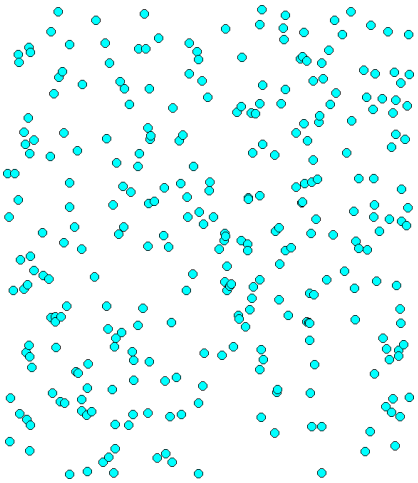


Abbildung D.8: Gephi 4,
300 Knoten

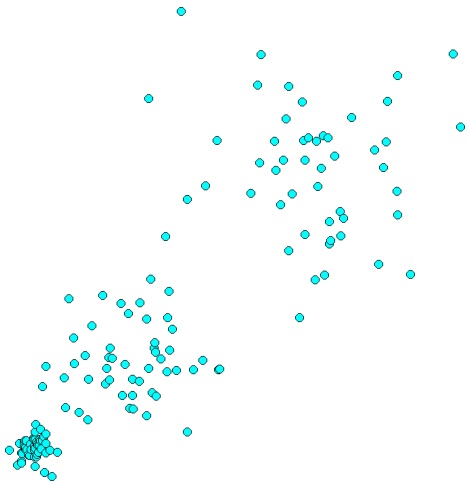


Abbildung D.9: Elki density,
150 Knoten

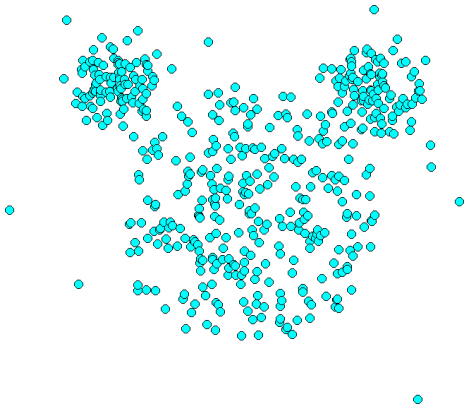


Abbildung D.10: Elki Mouse,
500 Knoten

D.2 Logistik

Die Abbildungen D.11 bis D.20 gehören zu den Szenarien der Logistik (Kartendaten © OpenStreetMap-Mitwirkende, CC-BY-SA).

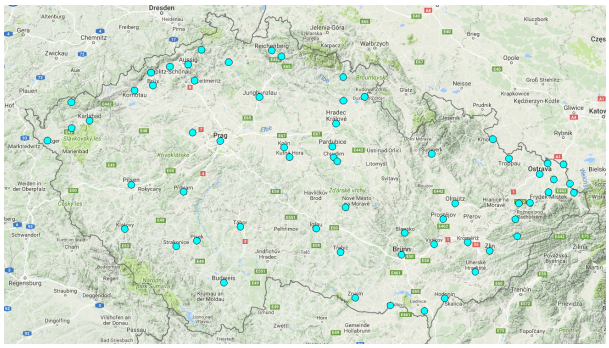


Abbildung D.11: Tschechien [315], 62 Knoten

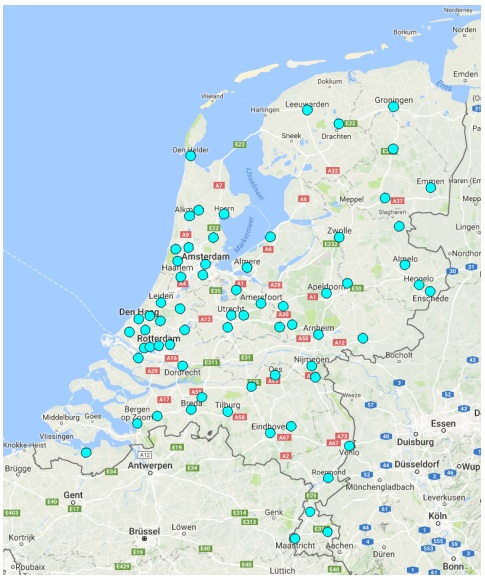


Abbildung D.12: Niederlande [316], 67 Knoten

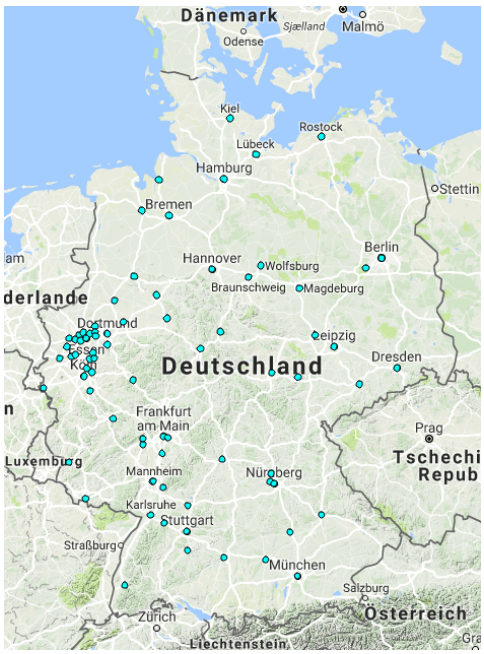


Abbildung D.13: Deutschland [317], 77 Knoten

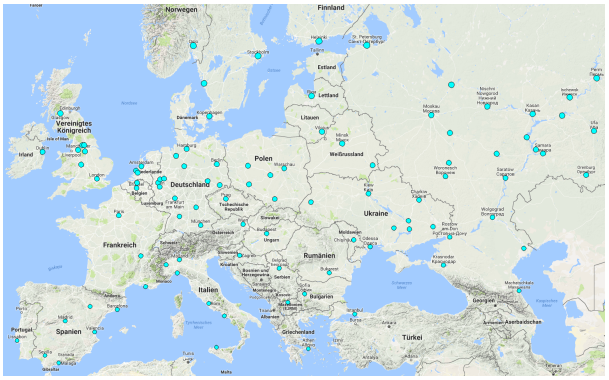


Abbildung D.14: Europa [318], 99 Knoten

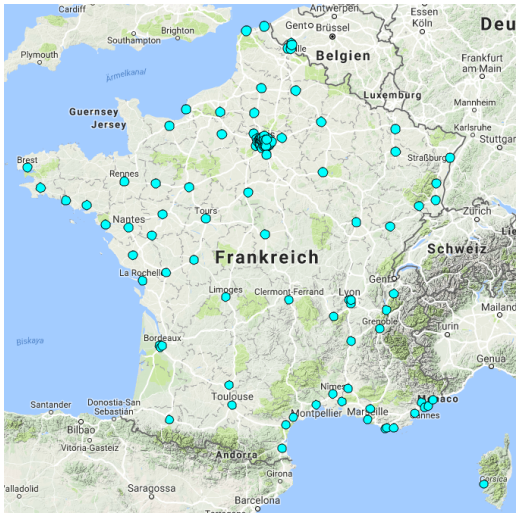


Abbildung D.15: Frankreich [319],
114 Knoten

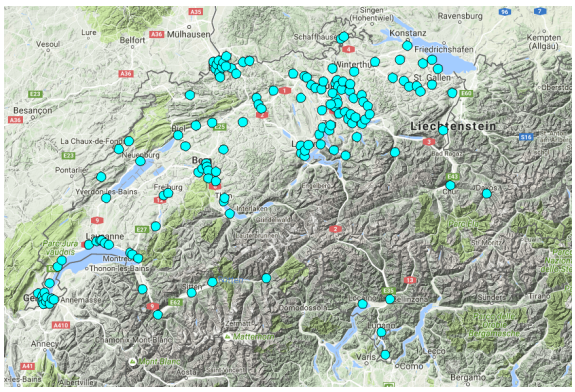


Abbildung D.16: Schweiz [320],
144 Knoten

Abbildung D.17: Italien [321],
147 Knoten

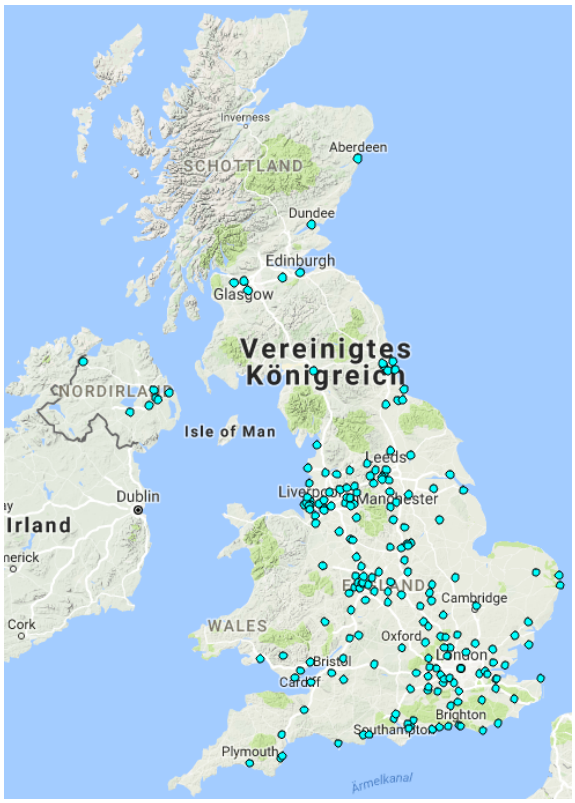


Abbildung D.18: Vereinigtes Königreich [322],
196 Knoten

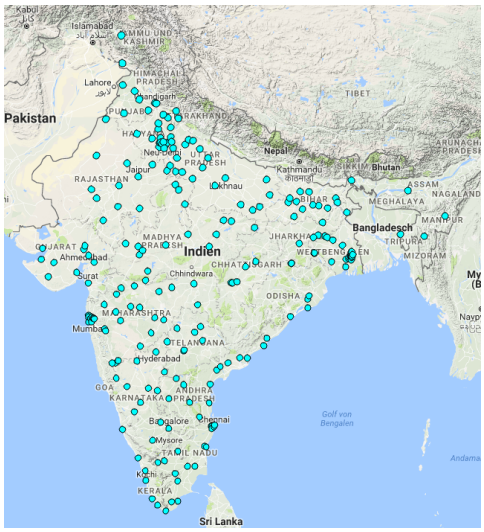


Abbildung D.19: Indien [323],
230 Knoten

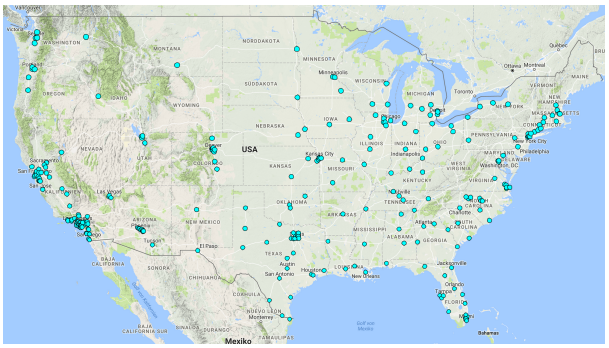


Abbildung D.20: United States of America [324],
273 Knoten

D.3 Rechnernetze

Die Abbildungen D.21 bis D.30 gehören zu den Szenarien der Rechnernetze (Datensatz von [19]).

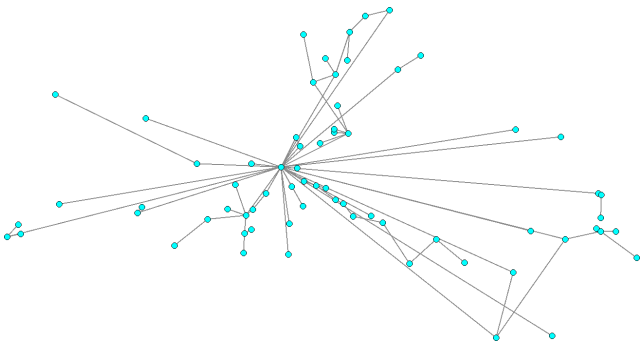


Abbildung D.21: Latnet,
68 Knoten und 73 Kanten

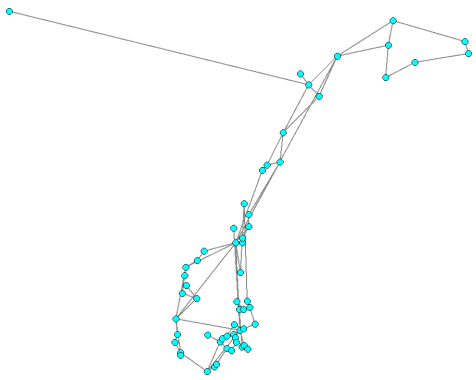


Abbildung D.22: Uninett,
47 Knoten und 102 Kanten

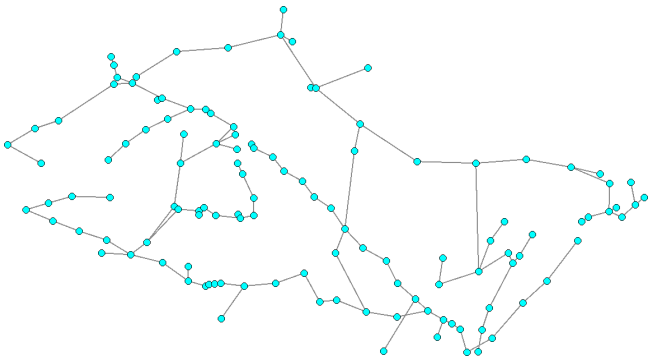


Abbildung D.23: DialtelecomCz,
124 Knoten und 126 Kanten

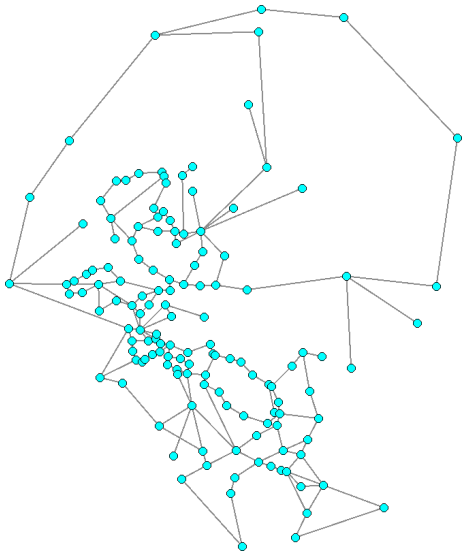


Abbildung D.24: GtsCe,
141 Knoten und 181 Kanten

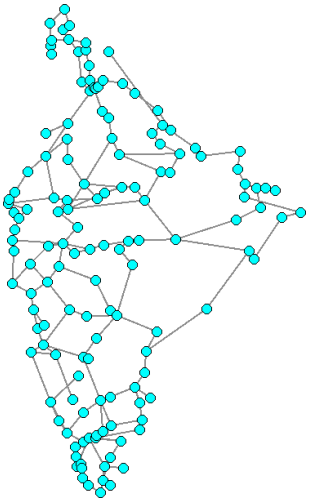


Abbildung D.25: TataNld,
143 Knoten und 181 Kanten

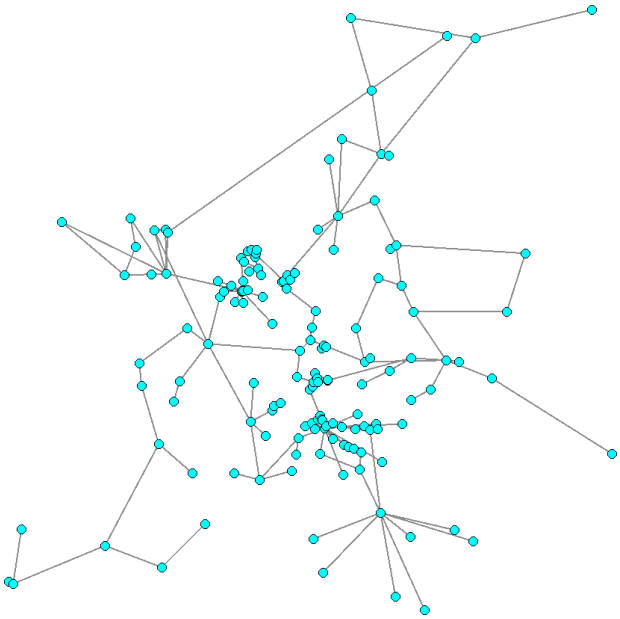


Abbildung D.26: Colt,
146 Knoten und 164 Kanten

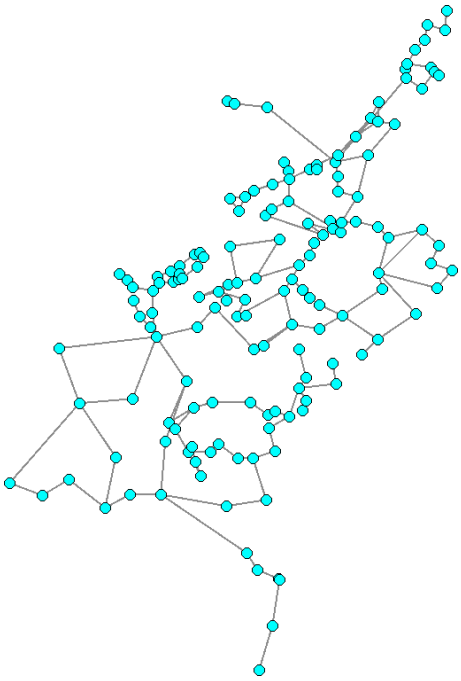


Abbildung D.27: US Carrier,
151 Knoten und 174 Kanten

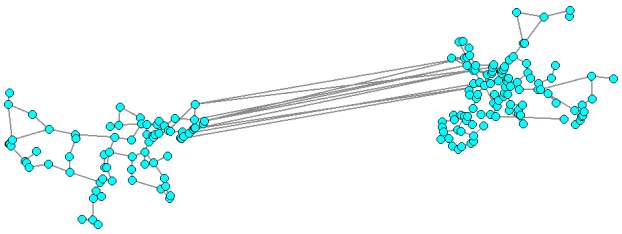


Abbildung D.28: Cogentco,
180 Knoten und 210 Kanten

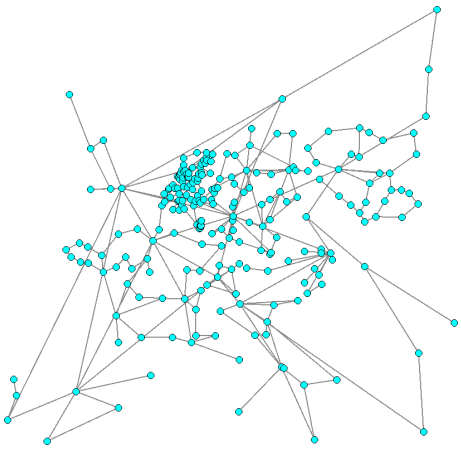


Abbildung D.29: Westeuropa,
271 Knoten und 385 Kanten

Abbildung D.30: Europa,
447 Knoten und 605 Kanten

E Evaluation

Dieser Abschnitt enthält weitere Analyseergebnisse zu den Algorithmen Lloyd und MacQueen. Es werden die Leistungsunterschiede bei den Nebenbedingungen Knotenplatzierung und freie Platzierung dargestellt.

E.1 Vergleich der Auswirkung von Platzierungsbedingungen von Lloyd

Die Abbildung E.1 stellt die Entfernungsunterschiede bei der jeweiligen Anzahl an Zentren dar. Die Tabelle E.1 zeigt die statistischen Werte zu dem Unterschied bei den Platzierungsbedingungen Knotenplatzierung und freie Platzierung.

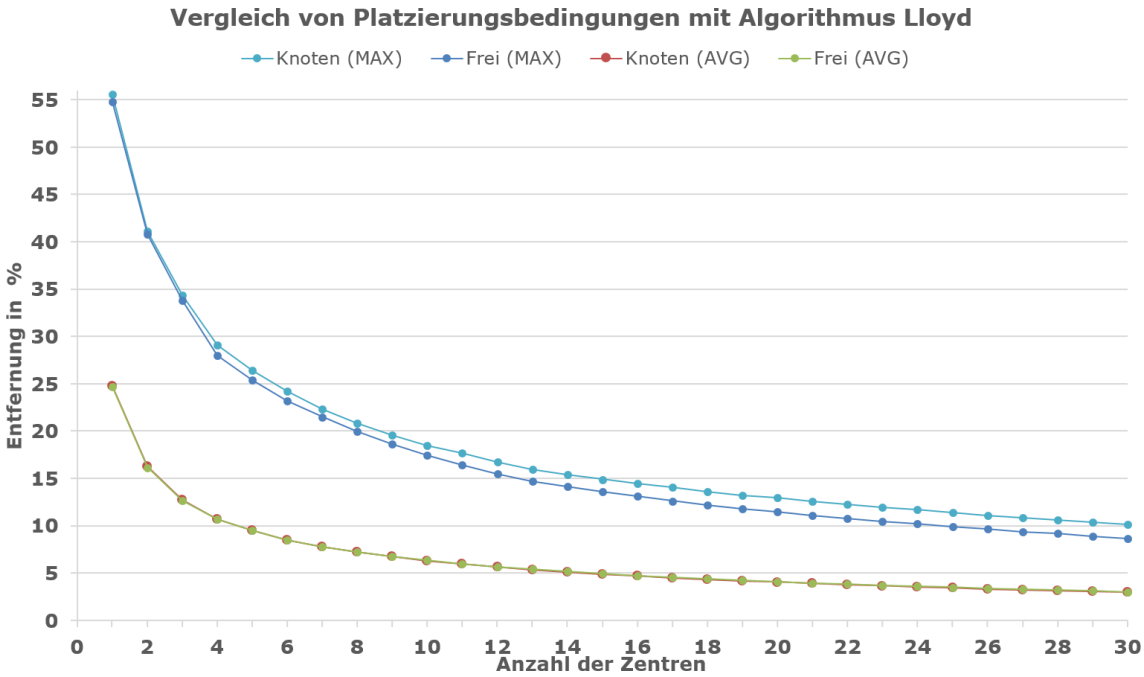


Abbildung E.1: Vergleich der Leistung von Llyod bei der Platzierung von Zentren unter Beachtung der Nebenbedingung knotenbasierte und freie Platzierung

Tabelle E.1: Statistische Auswertung der Leistungsdifferenzen von Llyod bei der Platzierung bei den Nebenbedingungen Knotenplatzierung und freie Platzierung

Kriterium	Durchschnitt	Median	Maximum	Minimum
Maximale Entfernung	8,5 %	9,1 %	14,8 %	0,8 %
Durchschnittliche Entfernung	-0,9 %	-0,8 %	0,5 %	-2,1 %

E.2 Vergleich der Auswirkung von Platzierungsbedingungen von MacQueen

Die Abbildung E.2 stellt die Entfernungsunterschiede bei der jeweiligen Anzahl an Zentren dar. Die Tabelle E.2 zeigt die statistischen Werte zu dem Unterschied bei den Platzierungsbedingungen Knotenplatzierung und freie Platzierung.

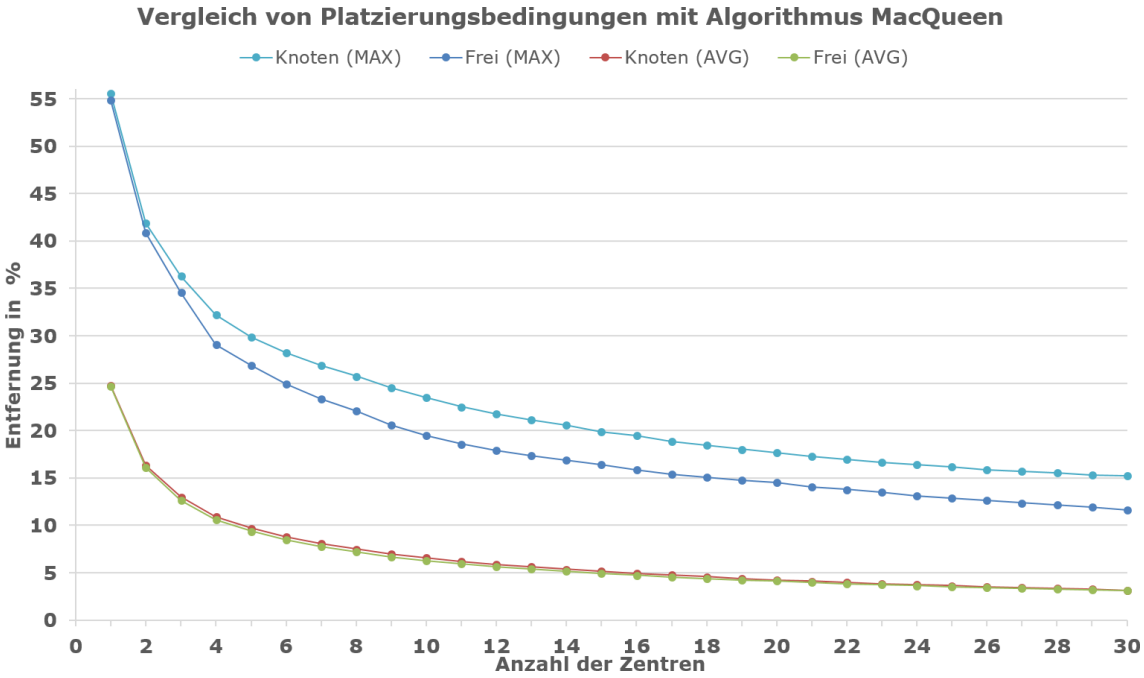


Abbildung E.2: Vergleich der Leistung von MacQueen bei der Platzierung von Zentren unter Beachtung der Nebenbedingung knotenbasierte und freie Platzierung

Tabelle E.2: Statistische Auswertung der Leistungsdifferenzen von MacQueen bei der Platzierung bei den Nebenbedingungen Knotenplatzierung und freie Platzierung

Kriterium	Durchschnitt	Median	Maximum	Minimum
Maximale Entfernung	16,2 %	18,0 %	23,9 %	1,4 %
Durchschnittliche Entfernung	3,3 %	3,2 %	4,6 %	0,3 %

E.3 Differenzen des allgemeinen Leistungsvergleichs

Die Tabelle E.3 zeigt die minimale und maximale Entfernung bei den vereinigten Testdaten von den verschiedenen Optimierungsverfahren bei der Platzierung von Zentren.

Tabelle E.3: Leistungsvergleich für das Zielkriterium Minimierung der maximalen Entfernung in % (Kleinere Werte sind besser.)

Anzahl der Zentren	*Monte Carlo*		*Zwei-Approx*		*ILP*		*0-Greedy*		*GRASP*		*MacQueen*		*Lloyd*		*k-Means++ (Frei)*		*GA*		*SA*		**Dragoon**		**Dragoon+**		**Dragoon (Frei)**		**Backtrack**	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
1	55,7	94,7	55,7	94,7	95,2	95,2	52,2	52,2	52,2	52,2	55,6	55,6	55,6	55,6	54,8	54,8	52,2	52,2	52,2	52,2	52,2	52,2	52,2	52,2	50,3	50,3	52,2	52,2
2	41,4	81,7	42,3	59,4	54,1	54,1	49,2	49,2	34,5	38,2	39,5	45,5	39,8	42,4	39,6	42,7	34,3	34,8	34,3	34,8	34,8	34,8	34,8	34,8	32,9	32,9	34,5	34,5
3	36,5	72,3	34,1	48,5	39,2	39,2	34,4	46,9	28,5	33,4	31,1	42,4	31,3	39,4	30,5	37,6	27,3	29,3	27,3	28,7	28,5	28,5	28,5	28,5	26,3	26,3	27,7	27,7
4	31,6	67,1	28,6	38,3	32,6	32,6	29,8	43,0	25,1	30,5	27,1	38,9	26,3	34,0	25,2	33,0	22,4	25,3	22,4	23,8	23,3	23,3	23,3	23,3	21,3	21,3	22,5	22,5
5	29,8	62,3	25,0	32,6	28,3	28,3	25,7	38,8	22,8	28,5	24,0	36,1	23,0	31,2	22,3	29,7	19,9	21,9	19,8	21,1	20,6	20,6	20,5	20,5	19,1	19,1	19,9	19,9
10	22,5	46,1	15,6	19,1	17,5	17,5	15,9	24,0	17,1	22,6	17,3	30,0	15,0	23,9	14,0	21,0	13,5	16,2	13,3	14,7	13,8	13,8	13,5	13,5	12,5	12,5	12,8	12,8
15	19,1	40,9	12,0	14,2	13,5	13,5	12,5	17,3	15,1	20,1	14,5	27,7	12,3	19,2	10,7	16,8	10,9	12,7	10,6	12,4	11,3	11,3	10,8	10,8	9,7	9,7	10,0	10,0
20	16,4	36,1	9,8	11,3	11,0	11,0	10,4	13,9	13,4	18,2	13,1	24,6	10,2	16,6	8,9	13,9	9,3	11,0	8,9	10,6	9,2	9,2	8,9	8,9	8,3	8,3	8,3	8,3
25	15,0	33,9	8,4	9,5	9,2	9,2	9,0	11,6	12,9	17,1	12,2	23,3	9,0	15,2	7,5	11,6	8,1	10,6	7,7	9,6	8,0	8,0	7,6	7,6	7,0	7,0	7,1	7,1
30	13,8	31,0	7,3	8,2	8,0	8,0	7,8	10,0	11,6	16,1	11,3	20,4	8,1	13,6	6,4	10,0	7,3	9,1	6,9	9,0	7,1	7,1	6,7	6,7	6,2	6,2	6,2	6,2
35	12,9	29,2	6,5	7,2	7,1	7,1	6,9	8,6	10,9	15,4	10,5	19,5	7,4	12,4	5,8	9,0	6,6	8,5	6,2	7,9	6,4	6,4	5,9	5,9	5,4	5,4	5,6	5,6
40	12,4	27,3	5,8	6,4	6,3	6,3	6,2	7,6	10,3	14,7	9,9	18,8	6,9	11,6	5,2	8,3	6,1	7,8	5,8	6,9	5,8	5,8	5,4	5,4	4,9	4,9	5,0	5,0
45	11,5	26,7	5,2	5,7	5,7	5,7	5,6	6,9	9,6	14,2	9,5	18,0	6,3	10,4	4,5	7,1	5,6	7,2	5,3	7,6	5,2	5,2	4,9	4,9	4,4	4,4	4,6	4,6
50	10,9	25,3	4,7	5,2	5,1	5,1	5,1	6,2	9,0	13,8	8,8	17,1	5,9	9,9	4,1	6,6	5,3	6,5	4,9	6,3	4,7	4,7	4,5	4,5	3,8	3,8	4,2	4,2

F Geolokalisation

Im Folgenden wird die Formel zur Schnittpunktberechnung von zwei Kreisen aufgeführt. Im Rahmen der Geolokalisation erfolgt zudem ein Vergleich der Abbildungsfunktionen bei der Verwendung unterschiedlicher Entfernungsbestimmungen. Anschließend sind die genutzten Daten zur Evaluation des Modells zur Geolokalisation aufgeführt.

F.1 Berechnung der Schnittpunkte von zwei Kreisen

Für die Lateration erfolgt die Berechnung der Schnittpunkte von zwei Kreisen. Dies geschieht ohne Winkelberechnung auf Grund der gekrümmten Form der Erde. Die quadratische Gleichung F.1 zeigt die allgemeine Lösung für die Schnittpunkte von zwei Kreisen aufgelöst nach dem y-Wert, wobei dieser der Gleichung F.2 genügt. Auf Grund der Übersichtlichkeit sind die Teilgleichungen F.3 bis F.6 separat aufgeführt.

$$\mathbf{Y'^2} - \frac{(a+b)*d}{c^2+d^2} * \mathbf{Y'} - R_1^2 - \left(\frac{a+b}{2*c}\right)^2 = 0 \tag{F.1}$$

$$Y = Y' + Y_1 \tag{F.2}$$

$$a = X_1^2 + Y_1^2 - R_1^2 - X_2^2 - Y_2^2 + R_2^2 \tag{F.3}$$

$$b = -2\,(X_1 - X_2) * X_1 - 2\,(Y_1 - Y_2) * Y_1 \tag{F.4}$$

$$c = (X_1 - X_2) \tag{F.5}$$

$$d = (Y_1 - Y_2) \tag{F.6}$$

F.2 Vergleich der Abbildungsfunktionen

Die Abbildungen F.1 und F.2 zeigen jeweils die gemittelte Abbildungsfunktion der VPs bei verschiedenen zu Grunde gelegten Verfahren zur Entfernungsbestimmung.

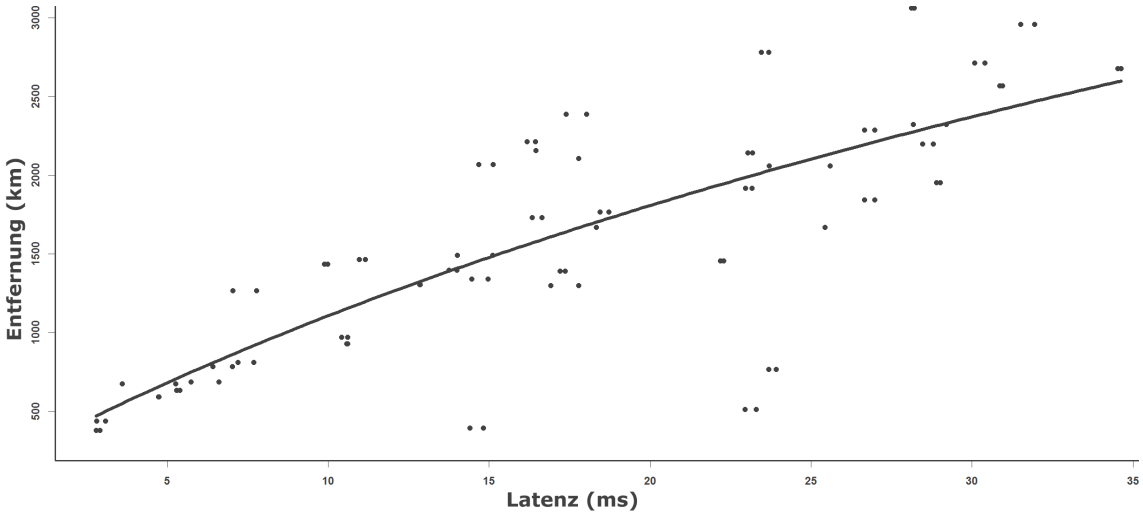


Abbildung F.1: Abbildungsfunktion gemäß der Korrelation von Latenz und Entfernungen nach Google Maps

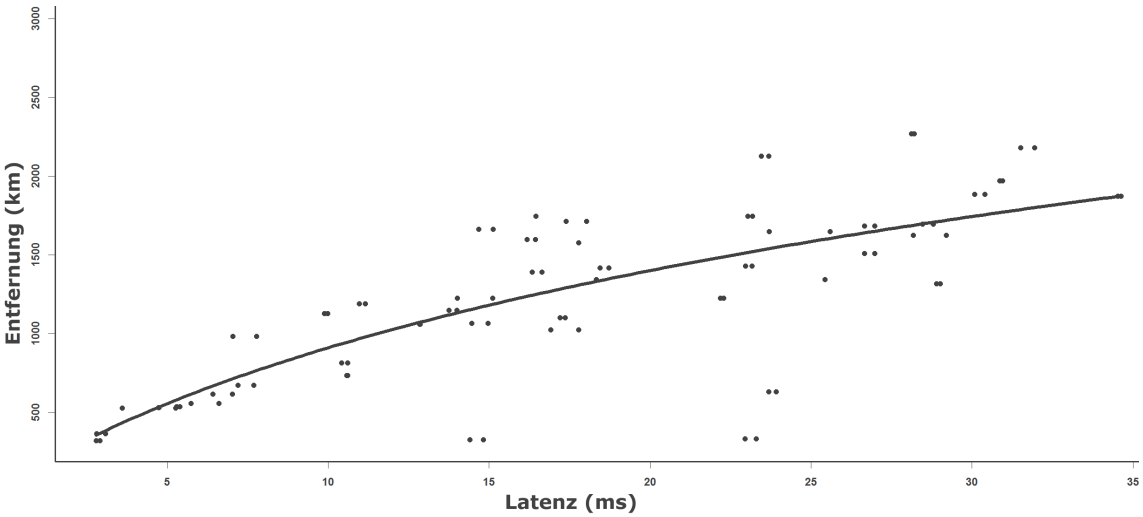


Abbildung F.2: Abbildungsfunktion gemäß der Korrelation von Latenz und orthodromen Entfernungen

F.3 Datensatz zur Geolokalisation

Die Tabelle F.1 zeigt die genutzten Daten zur Evaluation des spezifischen Modells der Geolokalisation (Stand 02.02.2016).

Tabelle F.1: Datensatz der verwendeten Ziele zur Überprüfung der IP Geolokalisation

Stadt	Land	Latitude	Longitude	ASN	ID	Name	IP-Adresse
Amsterdam	NL	52,345432	4,832429	12041	6082	nl-ams-as12041	199.10.66.231
Ballerup	DK	55,728092	12,377482	39839	6103	dk-blp-as39839	193.163.102.206
Berlin	DE	52,500122	13,401483	25291	6017	de-ber-as25291	37.49.152.204
Calw	DE	48,712210	8,747244	39702	6073	de-cal-as39702	89.106.219.238
Delft	NL	51,950000	4,225000	31019	6022	nl-dft-as31019	91.228.151.10
Ettlingen	DE	48,952507	8,390881	202040	6078	de-ett-as202040	193.141.27.220
Frankfurt	DE	50,120030	8,735270	5580	6023	de-fra-as5580	80.94.66.74
Glattbrugg	CH	47,435177	8,559989	20612	6115	ch-gtg-as20612	194.242.34.190
Göteborg	SE	57,729977	12,002948	50168	6063	se-got-as50168	185.47.192.253
Hamburg	DE	53,550000	10,048330	12731	6014	de-ham-as12731	213.128.137.33
Kiel	DE	54,366245	10,108079	13101	6052	de-kel-as13101	213.178.92.102
Prag	CZ	50,099430	14,392700	2852	6068	cz-prg-as2852	195.113.161.50
Stuttgart	DE	48,721450	9,128540	48918	6108	de-str-as48918	94.186.178.253
Turin	IT	45,091300	7,660600	12779	6011	it-trn-as12779	213.212.129.68
Wien	AT	48,208176	16,373820	1120	6042	at-vie-as1120	193.171.255.2
Zürich	CH	47,380123	8,545130	559	6015	ch-zrh-as559	130.59.80.2

G Vorveröffentlichungen

In diesem Abschnitt sind die wissenschaftlichen Veröffentlichungen und themenbezogenen Arbeiten aufgeführt, welche während der Arbeit an der Dissertation entstanden sind.

G.1 Vorveröffentlichungen mit Bezug zur Dissertation

2014

Titel:	A novel multi-agent system for complex scheduling problems
Autoren:	Peter Hillmann, Tobias Uhlig, Gabi Dreo Rodosek, Oliver Rose
Konferenz:	IEEE/ACM Winter Simulation Conference (WSC), Seite 231-241

2015

Titel:	On the Evaluation of Military Simulations: Towards a Taxonomy of Assessment Criteria
Autoren:	Mario Golling, Robert Koch, Peter Hillmann, Volker Eiseler, Lars Stiemert, Andres Rekker
Konferenz:	IEEE Military Communications and Information Systems Conference (MilCIS), Seiten 1-7

Titel:	Modelling of IP Geolocation by use of Latency Measurements
Autoren:	Peter Hillmann, Lars Stiemert, Gabi Dreo Rodosek, Oliver Rose
Konferenz:	IEEE Network and Service Management (CNSM), Seiten 173-177

Titel:	A Novel Approach to solve K-Center Problems with Geographical Placement
Autoren:	Peter Hillmann, Tobias Uhlig, Gabi Dreo Rodosek, Oliver Rose
Konferenz:	IEEE Service Operations and Logistics, and Informatics (SOLI), Seiten 31-36

Titel: An Efficient Solution for K-Center Problems
Autoren: Peter Hillmann, Tobias Uhlig, Gabi Dreo Rodosek, Oliver Rose
Konferenz: IEEE/ACM Winter Simulation Conference (WSC), Seiten 3186-3187

Titel: Dragoon: Advanced Modelling of IP Geolocation by use of Latency Measurements
Autoren: Peter Hillmann, Lars Stiemert, Gabi Dreo Rodosek, Oliver Rose
Konferenz: IEEE Internet Technology and Secured Transactions (ICITST), Seiten 438-445

2016

Titel: On the Path to High Precise IP Geolocation: A Self-Optimizing Model
Autoren: Peter Hillmann, Lars Stiemert, Gabi Dreo Rodosek, Oliver Rose
Journal: International Journal of Intelligent Computing Research (IJICR), eingeladen, Seiten 682-693

Titel: Simulation and Optimization of Content Delivery Networks Considering User Profiles and Preferences of Internet Service Providers
Autoren: Peter Hillmann, Tobias Uhlig, Gabi Dreo Rodosek, Oliver Rose
Konferenz: IEEE/ACM Winter Simulation Conference (WSC), Seiten 3143-3154

Titel: Evaluation of the General Applicability of Dragoon for the k-Center Problem
Autoren: Tobias Uhlig, Peter Hillmann, Oliver Rose
Konferenz: IEEE/ACM Winter Simulation Conference (WSC), Seiten 2284-2291

Titel: Modeling the Location Selection of Mirror Servers in Content Delivery Networks
Autoren: Peter Hillmann, Tobias Uhlig, Gabi Dreo Rodosek, Oliver Rose
Konferenz: IEEE Services Computing Conference/Bigdata Congress (ICWS/SCC/CLOUD/MS/Bigdata Congress/SERVICES), Seiten 438-445

2017

Titel:	Optimized Travel to Meetings on a Common Location of Geographical Distributed Participants
Autoren:	Peter Hillmann, Bastian Kühnel, Tobias Uhlig, Gabi Dreo Rodosek, and Oliver Rose
Konferenz:	IEEE Service Operations and Logistics, and Informatics (SOLI), Seiten 1-5

G.2 Übersicht der Vorveröffentlichungen zu Teilaspekten der Dissertation

Die Abbildung G.1 zeigt die einzelnen Anteile dieser Arbeit, welche in wissenschaftlichen Arbeiten vorab auf Konferenzen veröffentlicht wurden. In den Konferenzbeiträgen erfolgte eine Beschränkung auf die hauptsächlichen Aspekte ohne eine umfassende und tiefgehende Beschreibung. Die in der Abbildung aufgezeigten Bestandteile geben auf Grund des gesamtheitlichen Bezugs eine Übersicht, wobei trotz der farblichen Markierung Teilaspekte noch nicht veröffentlicht wurden. Insbesondere ist der vollständige Zusammenhang des generischen Ansatzes dieser Arbeit vorbehalten.

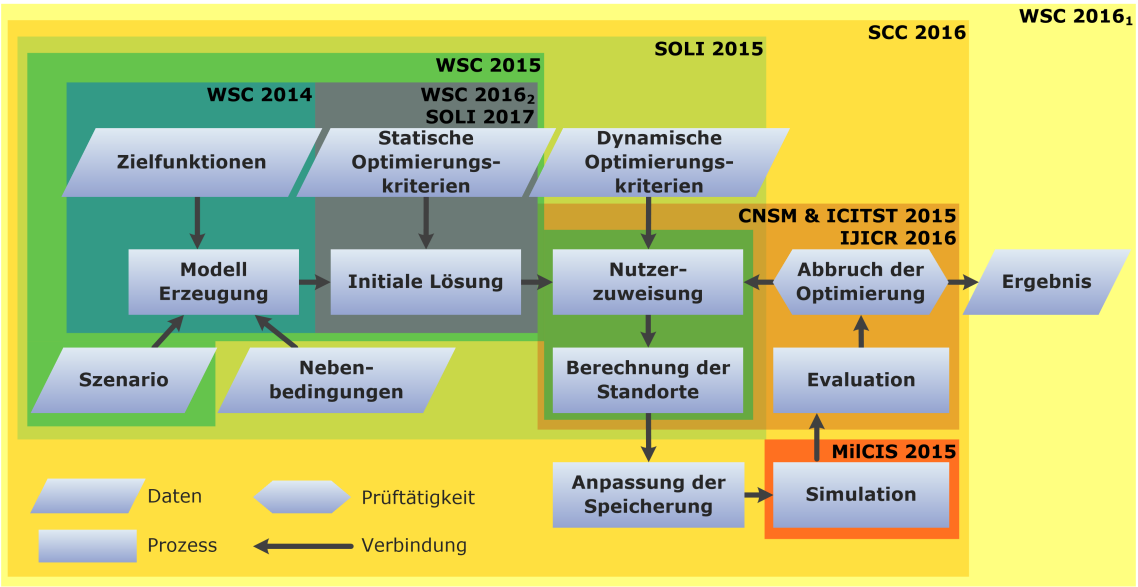


Abbildung G.1: Übersicht der Vorabveröffentlichungen und Einordnung in das Gesamtkonzept der Arbeit

G.3 Veröffentlichungen während der Arbeit an der Dissertation ohne direkten Bezug

2013

Titel:	Ganzheitliche Architektur zur Entwicklung und Analyse sicherheitskritischer Systeme und Anwendungen.
Autoren:	Mario Golling, Robert Koch, Peter Hillmann, Gabi Dreo Rodosek.
Konferenz:	6. DFN-Forum Kommunikationstechnologien, Seite 87-96

Titel:	A Flexible Cryptographic Infrastructure for High-security SDR-based Systems
Autoren:	Peter Hillmann, Björn Stelte
Konferenz:	IEEE Military Communications and Information Systems Conference (MilCIS), Seiten 1-6

2014

Titel:	YANG2UML: Bijective transformation and simplification of YANG to UML.
Autoren:	Mario Golling, Robert Koch, Peter Hillmann, Rick Hofstede, Frank Tietze
Konferenz:	IEEE Network and Service Management (CNSM), Seite 300-303

2015

Titel:	Strategie zur Verfolgung einzelner IP-Pakete zur Datenflussanalyse
Autoren:	Frank Tietze, Peter Hillmann, Gabi Dreo Rodosek
Konferenz:	8. DFN-Forum Kommunikationstechnologien, Anwärter für Auszeichnung als bestes Paper, Seiten 47-56

Titel:	Tracemax: A Novel Single Packet IP Traceback Strategy for Data-Flow Analysis
Autoren:	Peter Hillmann, Frank Tietze, Gabi Dreo Rodosek
Konferenz:	IEEE Local Computer Networks (LCN), Seiten 177-180

Titel:	Strategies for Tracking Individual IP Packets Towards DDoS
Autoren:	Peter Hillmann, Frank Tietze, Gabi Dreo Rodosek
Konferenz:	Praxis der Informationsverarbeitung und Kommunikation (PIK), eingeladen, Seiten 15-21

2016

Titel:	A Traffic Merging and Generation Framework for Realistic Synthesis of Network Traffic
Autoren:	Cem Gündogan, Sandro Passarelli, Peter Hillmann, Christian Dietz, Lars Stiemert
Konferenz:	SPRING Workshop, Seiten 1-2

2017

Titel:	CAKE: Hybrides Gruppen-Schlüssel-Management Verfahren
Autoren:	Peter Hillmann, Marcus Knüpfer und Gabi Dreo Rodosek
Konferenz:	10. DFN-Forum Kommunikationstechnologien, X-WiNner-Award für bestes Paper, Seiten 31-40

Erklärung

Ich versichere an Eides statt, dass ich die der Fakultät für Informatik
der Universität der Bundeswehr München vorgelegte Dissertation mit dem Thema

„Generische Modellierung von zweistufigen Transportketten zur Planung und Optimierung von Verteilzentren in komplexen Systemen“

ohne fremde Hilfe erstellt, bei der Abfassung keine anderen als die im Schriftenverzeichnis angeführten Hilfsmittel benutzt und die wissenschaftlichen Leistungen eigenständig erbracht habe.

Die Dissertation wurde betreut von Prof. Dr. Gabi Dreo Rodosek
Prof. Dr. Oliver Rose

Ich habe die Dissertation noch nicht veröffentlicht.

Ich habe an keiner in- oder ausländischen wissenschaftlichen Hochschule ein Gesuch um Zulassung zur Promotion eingereicht oder die vorliegende oder eine ähnliche Arbeit als Dissertation vorgelegt.

Ich habe am 24.06.15 in der Universität der Bundeswehr München
bei der Fakultät für Informatik
unter Vorlage einer Dissertation mit dem Thema (Arbeitstitel)

„Entscheidungsunterstützungssysteme für Distributionszentren und Zwischenspeicher von Versorgungsnetzen“

die Zulassung zur Promotion beantragt.
Das Ergebnis lautet: Bestätigung der Aufnahme in die Promotionsliste

Die Promotionsordnung der Universität der Bundeswehr München ist mir bekannt.

Neubiberg, den 24. November 2017

Peter Hillmann
.....
(Unterschrift)