清华大学优秀博士学位论文丛书

# 供应链中的资源分配研究
## ——以新能源汽车与数据中心为例

唐润宇（Tang Runyu）著

Resource Allocation in Supply Chain:
Analytics for NEV and Data Center Supply Chain

清华大学出版社
北京

## 内 容 简 介

新兴技术给供应链管理者带来了机遇与挑战，本书主要聚焦于新兴技术下的供应链中存在的资源分配问题，主要以新能源汽车与数据中心供应链为例，分别研究了新能源汽车市场的政府财政补贴资源分配问题、数据中心供应链的网络设计和服务资源分配问题以及数据中心供应链中在服务水平协议下的云计算服务资源的供给和分配问题。本书综合使用了管理科学与工程领域的多种工具，包括博弈论、整数规划、鲁棒优化等方法，能够为决策者提供有效的辅助决策工具和丰富的管理启示。



**图书在版编目（CIP）数据**

供应链中的资源分配研究：以新能源汽车与数据中心为例 / 唐润宇著. —北京：清华大学出版社，2024.5

（清华大学优秀博士学位论文丛书）

ISBN 978-7-302-65887-0

Ⅰ. ①供… Ⅱ. ①唐… Ⅲ. ①新能源–汽车工业–供应链管理 Ⅳ. ①F407.471

中国国家版本馆 CIP 数据核字 (2024) 第 065103 号

**责任编辑：**张维嘉
**封面设计：**傅瑞学
**责任校对：**赵丽敏
**责任印制：**刘海龙

**出版发行：**清华大学出版社
网　　址：https://www.tup.com.cn，https://www.wqxuetang.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-83470000　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
**印 装 者：**三河市东方印刷有限公司
**经　　销：**全国新华书店
**开　　本：**155mm×235mm　　**印　张：**12.25　　**插　　页：**1　　**字　　数：**209 千字
**版　　次：**2024 年 5 月第 1 版　　**印　　次：**2024 年 5 月第 1 次印刷
**定　　价：**99.00 元

产品编号：101487-01

# 一流博士生教育
# 体现一流大学人才培养的高度（代丛书序）[①]

人才培养是大学的根本任务。只有培养出一流人才的高校，才能够成为世界一流大学。本科教育是培养一流人才最重要的基础，是一流大学的底色，体现了学校的传统和特色。博士生教育是学历教育的最高层次，体现出一所大学人才培养的高度，代表着一个国家的人才培养水平。清华大学正在全面推进综合改革，深化教育教学改革，探索建立完善的博士生选拔培养机制，不断提升博士生培养质量。

## 学术精神的培养是博士生教育的根本

学术精神是大学精神的重要组成部分，是学者与学术群体在学术活动中坚守的价值准则。大学对学术精神的追求，反映了一所大学对学术的重视、对真理的热爱和对功利性目标的摒弃。博士生教育要培养有志于追求学术的人，其根本在于学术精神的培养。

无论古今中外，博士这一称号都和学问、学术紧密联系在一起，和知识探索密切相关。我国的博士一词起源于 2000 多年前的战国时期，是一种学官名。博士任职者负责保管文献档案、编撰著述，须知识渊博并负有传授学问的职责。东汉学者应劭在《汉官仪》中写道："博者，通博古今；士者，辩于然否。"后来，人们逐渐把精通某种职业的专门人才称为博士。博士作为一种学位，最早产生于 12 世纪，最初它是加入教师行会的一种资格证书。19 世纪初，德国柏林大学成立，其哲学院取代了以往神学院在大学中的地位，在大学发展的历史上首次产生了由哲学院授予的哲学博士学位，并赋予了哲学博士深层次的教育内涵，即推崇学术自由、创造新知识。哲学博士的设立标志着现代博士生教育的开端，博士则被定义为

① 本文首发于《光明日报》，2017 年 12 月 5 日。

独立从事学术研究、具备创造新知识能力的人，是学术精神的传承者和光大者。

博士生学习期间是培养学术精神最重要的阶段。博士生需要接受严谨的学术训练，开展深入的学术研究，并通过发表学术论文、参与学术活动及博士论文答辩等环节，证明自身的学术能力。更重要的是，博士生要培养学术志趣，把对学术的热爱融入生命之中，把捍卫真理作为毕生的追求。博士生更要学会如何面对干扰和诱惑，远离功利，保持安静、从容的心态。学术精神，特别是其中所蕴含的科学理性精神、学术奉献精神，不仅对博士生未来的学术事业至关重要，对博士生一生的发展都大有裨益。

## 独创性和批判性思维是博士生最重要的素质

博士生需要具备很多素质，包括逻辑推理、言语表达、沟通协作等，但是最重要的素质是独创性和批判性思维。

学术重视传承，但更看重突破和创新。博士生作为学术事业的后备力量，要立志于追求独创性。独创意味着独立和创造，没有独立精神，往往很难产生创造性的成果。1929 年 6 月 3 日，在清华大学国学院导师王国维逝世二周年之际，国学院师生为纪念这位杰出的学者，募款修造“海宁王静安先生纪念碑”，同为国学院导师的陈寅恪先生撰写了碑铭，其中写道：“先生之著述，或有时而不章；先生之学说，或有时而可商；惟此独立之精神，自由之思想，历千万祀，与天壤而同久，共三光而永光。”这是对于一位学者的极高评价。中国著名的史学家、文学家司马迁所讲的“究天人之际，通古今之变，成一家之言”也是强调要在古今贯通中形成自己独立的见解，并努力达到新的高度。博士生应该以“独立之精神、自由之思想”来要求自己，不断创造新的学术成果。

诺贝尔物理学奖获得者杨振宁先生曾在 20 世纪 80 年代初对到访纽约州立大学石溪分校的 90 多名中国学生、学者提出：“独创性是科学工作者最重要的素质。”杨先生主张做研究的人一定要有独创的精神、独到的见解和独立研究的能力。在科技如此发达的今天，学术上的独创性变得越来越难，也愈加珍贵和重要。博士生要树立敢为天下先的志向，在独创性上下功夫，勇于挑战最前沿的科学问题。

批判性思维是一种遵循逻辑规则、不断质疑和反省的思维方式，具有批判性思维的人勇于挑战自己，敢于挑战权威。批判性思维的缺乏往往被认为是中国学生特有的弱项，也是我们在博士生培养方面存在的一

个普遍问题。2001 年，美国卡内基基金会开展了一项“卡内基博士生教育创新计划”，针对博士生教育进行调研，并发布了研究报告。该报告指出：在美国和欧洲，培养学生保持批判而质疑的眼光看待自己、同行和导师的观点同样非常不容易，批判性思维的培养必须成为博士生培养项目的组成部分。

对于博士生而言，批判性思维的养成要从如何面对权威开始。为了鼓励学生质疑学术权威、挑战现有学术范式，培养学生的挑战精神和创新能力，清华大学在 2013 年发起“巅峰对话”，由学生自主邀请各学科领域具有国际影响力的学术大师与清华学生同台对话。该活动迄今已经举办了 21 期，先后邀请 17 位诺贝尔奖、3 位图灵奖、1 位菲尔兹奖获得者参与对话。诺贝尔化学奖得主巴里·夏普莱斯（Barry Sharpless）在 2013 年 11 月来清华参加“巅峰对话”时，对于清华学生的质疑精神印象深刻。他在接受媒体采访时谈道：“清华的学生无所畏惧，请原谅我的措辞，但他们真的很有胆量。”这是我听到的对清华学生的最高评价，博士生就应该具备这样的勇气和能力。培养批判性思维更难的一层是要有勇气不断否定自己，有一种不断超越自己的精神。爱因斯坦说：“在真理的认识方面，任何以权威自居的人，必将在上帝的嬉笑中垮台。”这句名言应该成为每一位从事学术研究的博士生的箴言。

## 提高博士生培养质量有赖于构建全方位的博士生教育体系

一流的博士生教育要有一流的教育理念，需要构建全方位的教育体系，把教育理念落实到博士生培养的各个环节中。

在博士生选拔方面，不能简单按考分录取，而是要侧重评价学术志趣和创新潜力。知识结构固然重要，但学术志趣和创新潜力更关键，考分不能完全反映学生的学术潜质。清华大学在经过多年试点探索的基础上，于 2016 年开始全面实行博士生招生“申请–审核”制，从原来的按照考试分数招收博士生，转变为按科研创新能力、专业学术潜质招收，并给予院系、学科、导师更大的自主权。《清华大学“申请–审核”制实施办法》明晰了导师和院系在考核、遴选和推荐上的权力和职责，同时确定了规范的流程及监管要求。

在博士生指导教师资格确认方面，不能论资排辈，要更看重教师的学术活力及研究工作的前沿性。博士生教育质量的提升关键在于教师，要让更多、更优秀的教师参与到博士生教育中来。清华大学从 2009 年开始探

索将博士生导师评定权下放到各学位评定分委员会，允许评聘一部分优秀副教授担任博士生导师。近年来，学校在推进教师人事制度改革过程中，明确教研系列助理教授可以独立指导博士生，让富有创造活力的青年教师指导优秀的青年学生，师生相互促进、共同成长。

在促进博士生交流方面，要努力突破学科领域的界限，注重搭建跨学科的平台。跨学科交流是激发博士生学术创造力的重要途径，博士生要努力提升在交叉学科领域开展科研工作的能力。清华大学于 2014 年创办了“微沙龙”平台，同学们可以通过微信平台随时发布学术话题，寻觅学术伙伴。3 年来，博士生参与和发起“微沙龙”12 000 多场，参与博士生达 38 000 多人次。“微沙龙”促进了不同学科学生之间的思想碰撞，激发了同学们的学术志趣。清华于 2002 年创办了博士生论坛，论坛由同学自己组织，师生共同参与。博士生论坛持续举办了 500 期，开展了 18 000 多场学术报告，切实起到了师生互动、教学相长、学科交融、促进交流的作用。学校积极资助博士生到世界一流大学开展交流与合作研究，超过 60%的博士生有海外访学经历。清华于 2011 年设立了发展中国家博士生项目，鼓励学生到发展中国家亲身体验和调研，在全球化背景下研究发展中国家的各类问题。

在博士学位评定方面，权力要进一步下放，学术判断应该由各领域的学者来负责。院系二级学术单位应该在评定博士论文水平上拥有更多的权力，也应担负更多的责任。清华大学从 2015 年开始把学位论文的评审职责授权给各学位评定分委员会，学位论文质量和学位评审过程主要由各学位分委员会进行把关，校学位委员会负责学位管理整体工作，负责制度建设和争议事项处理。

全面提高人才培养能力是建设世界一流大学的核心。博士生培养质量的提升是大学办学质量提升的重要标志。我们要高度重视、充分发挥博士生教育的战略性、引领性作用，面向世界、勇于进取，树立自信、保持特色，不断推动一流大学的人才培养迈向新的高度。

清华大学校长

2017 年 12 月

# 丛书序二

以学术型人才培养为主的博士生教育，肩负着培养具有国际竞争力的高层次学术创新人才的重任，是国家发展战略的重要组成部分，是清华大学人才培养的重中之重。

作为首批设立研究生院的高校，清华大学自 20 世纪 80 年代初开始，立足国家和社会需要，结合校内实际情况，不断推动博士生教育改革。为了提供适宜博士生成长的学术环境，我校一方面不断地营造浓厚的学术氛围，一方面大力推动培养模式创新探索。我校从多年前就已开始运行一系列博士生培养专项基金和特色项目，激励博士生潜心学术、锐意创新，拓宽博士生的国际视野，倡导跨学科研究与交流，不断提升博士生培养质量。

博士生是最具创造力的学术研究新生力量，思维活跃，求真求实。他们在导师的指导下进入本领域研究前沿，吸取本领域最新的研究成果，拓宽人类的认知边界，不断取得创新性成果。这套优秀博士学位论文丛书，不仅是我校博士生研究工作前沿成果的体现，也是我校博士生学术精神传承和光大的体现。

这套丛书的每一篇论文均来自学校新近每年评选的校级优秀博士学位论文。为了鼓励创新，激励优秀的博士生脱颖而出，同时激励导师悉心指导，我校评选校级优秀博士学位论文已有 20 多年。评选出的优秀博士学位论文代表了我校各学科最优秀的博士学位论文的水平。为了传播优秀的博士学位论文成果，更好地推动学术交流与学科建设，促进博士生未来发展和成长，清华大学研究生院与清华大学出版社合作出版这些优秀的博士学位论文。

感谢清华大学出版社，悉心地为每位作者提供专业、细致的写作和出

版指导，使这些博士论文以专著方式呈现在读者面前，促进了这些最新的优秀研究成果的快速广泛传播。相信本套丛书的出版可以为国内外各相关领域或交叉领域的在读研究生和科研人员提供有益的参考，为相关学科领域的发展和优秀科研成果的转化起到积极的推动作用。

感谢丛书作者的导师们。这些优秀的博士学位论文，从选题、研究到成文，离不开导师的精心指导。我校优秀的师生导学传统，成就了一项项优秀的研究成果，成就了一大批青年学者，也成就了清华的学术研究。感谢导师们为每篇论文精心撰写序言，帮助读者更好地理解论文。

感谢丛书的作者们。他们优秀的学术成果，连同鲜活的思想、创新的精神、严谨的学风，都为致力于学术研究的后来者树立了榜样。他们本着精益求精的精神，对论文进行了细致的修改完善，使之在具备科学性、前沿性的同时，更具系统性和可读性。

这套丛书涵盖清华众多学科，从论文的选题能够感受到作者们积极参与国家重大战略、社会发展问题、新兴产业创新等的研究热情，能够感受到作者们的国际视野和人文情怀。相信这些年轻作者们勇于承担学术创新重任的社会责任感能够感染和带动越来越多的博士生，将论文书写在祖国的大地上。

祝愿丛书的作者们、读者们和所有从事学术研究的同行们在未来的道路上坚持梦想，百折不挠！在服务国家、奉献社会和造福人类的事业中不断创新，做新时代的引领者。

相信每一位读者在阅读这一本本学术著作的时候，在吸取学术创新成果、享受学术之美的同时，能够将其中所蕴含的科学理性精神和学术奉献精神传播和发扬出去。

清华大学研究生院院长

2018 年 1 月 5 日

# 导师序言

在这个快速发展的技术时代，资源的有效分配已经成为供应链管理的核心问题。本书旨在深入探讨供应链中资源分配的动态和策略，特别是聚焦于新兴的新能源汽车行业和不断增长的数据中心领域。随着环境问题的日益严峻和技术的快速进步，这两个领域对资源分配的需求和挑战提供了一个独特且紧迫的研究视角。

首先，本书探讨了新能源汽车行业的供应链的补贴政策。考虑到消费者对电动汽车的里程焦虑，政府在发放补贴促进电动汽车推广时，可以直接提供补贴给消费者，也可以扶持充电桩建设从而间接提高电动汽车的吸引力。如何分配补贴资源能够更快更高效地推广电动汽车并提升社会福利是一个非常值得研究的问题。本书采用的模型很好地构建了政府、消费者和充电桩投资人的三方博弈关系，在求解出最优补贴政策的同时也得到了丰富的管理启示。

其次，本书研究了数据中心的资源分配问题。数据中心作为支撑当今数字经济的重要基础设施，其能源消耗和效率成为全球关注的焦点。本书具体研究了数据中心的网络布局问题以及数据中心内部的服务资源动态供给问题。

数据中心的选址与传统供应链的选址有着相似之处，但是由于其电能消耗成本较高，存在数据处理延迟等特性，在网络布局设计上也有着自己的特色。本书着重挖掘了数据中心内部的网络延迟成本给网络特征带来的影响，将网络布局问题建模成混合整数二阶锥优化问题，并提出了相关算法对问题进行求解，最后也通过使用实际数据进行敏感性分析得到了数据中心网络设计的一些重要性质。

对于数据中心内部的服务资源分配问题，本书着重考虑了如何在服

务水平协议下动态调节备用服务器资源，从而在保证服务质量的同时降低运营成本。为了解决服务器可能出现的随机宕机事件，本书使用了分布式鲁棒优化的框架来求解动态规划问题，提出了可行的算法，并通过数值实验验证了此框架下的服务资源供给方案确实能够相比已有文献带来更低的运营成本。

通过对这两个行业的深入分析，本书展示了供应链中资源分配的复杂性和多维度挑战。本书结合理论研究和实际数值案例，探讨有效的资源分配策略和方法。不仅为学术界提供新的研究方向，贡献新的管理理论，也为实践界的决策者和管理者在资源优化和供应链管理方面提供指导。随着全球经济和环境挑战的不断演变，供应链中资源分配的研究将变得越来越重要，希望本书能为供应链管理领域贡献新的见解和解决方案。

梁　湧

2024 年 1 月

# 摘　要

新技术的发展改变了人类的生活习惯，也改变了商业环境。新兴技术下的供应链管理成为商业实践亟待解决的问题，也为学术研究带来了新的挑战。本书以新能源汽车供应链及数据中心供应链这两个具有代表性的供应链为例，研究了供应链中面临的资源分配等运营管理问题。

首先，针对新能源汽车供应链，本书聚焦于电动汽车的补贴资源分配问题。电动汽车有着减少温室气体排放的优点，但是作为新兴技术并未被广泛接受，因此各地政府都在通过补贴政策促进电动汽车市场的发展。政府既可向购买电动汽车的消费者提供补贴，也可对充电桩的建设提供补贴。在电动汽车的推广过程中，充电桩和电动汽车作为互补产品形成了正反馈效应。综合考虑电动汽车对社会发展的益处和补贴支付的成本后，本书研究了政府能提供的最佳补贴资源分配策略，并研究了模型参数对补贴金额以及市场上电动汽车和充电桩数量的影响。最后还搜集了深圳市的相关数据，发现综合使用两种补贴较仅补贴消费者效果更佳。

其次，针对数据中心供应链，本书聚焦于数据中心的网络设计、需求分配和资源供给问题。数据中心是互联网服务的物理基础，需要很高的资本投入，不仅影响着企业短期运营层面的策略，也可能影响其长期战略层面的决策。因此，本书构建了一个整合优化模型，综合考虑了电能消耗成本、固定成本、延迟成本等，将问题构建成一个混合整数非线性规划问题。为了加快模型的求解速度，根据问题的具体结构性质构造了两种基于拉格朗日松弛方法的算法，并通过数值计算验证了算法的有效性。最后采集了相关实际数据，应用该模型设计出了数据中心网络，并通过参数变化的敏感性分析得到了数据中心建设的管理启示。

最后，针对数据中心供应链，本书还研究了服务水平协议下的云计算

服务资源分配。在服务水平协议中，服务供给方需要在合同期内保证其服务质量，否则需要给予消费方一定补偿。得益于物联网和互联网技术的发展，云计算服务供给方能够实时监控和调整其服务资源。因此本部分构造了一个鲁棒动态规划模型优化其动态资源调整策略。书中使用数据驱动的方法构建不确定集合，并设计了凸化算法，将每期的价值函数转化为分段线性凸函数，从而将鲁棒动态规划模型转化为若干个有限维的线性规划进行求解。最后搜集和生成了宕机时间的历史数据，应用提出的鲁棒动态规划模型，发现本书提出的算法确实可以显著降低成本。

**关键词**：新能源汽车供应链；数据中心供应链；资源分配；整数规划；鲁棒优化

# Abstract

The development of new technologies has reshaped both human lives and the business environment. The new supply chain management induced by emerging technologies becomes crucial to both business practice and academic researches. This book takes new energy vehicle supply chain and data center supply chain as examples to study resource allocation problems in new supply chain.

First, for new energy vehicle supply chain, this book focuses on the subsidy allocation strategies for electric vehicle (EV) adoption. The electric vehicles can reduce greenhouse gas emissions, but lack of adoption because of their novelty. Therefore, many governments provide incentives to promote EV adoption, including subsidizing EV buyers and charging station investors. By constructing an analytical model, which incorporate the positive network effect between EV and charging stations to derive the optimal subsidy policy, rich managerial insights are generated by studying on how model parameters affect the model outcomes. Through collecting real-world data from Shenzhen, it comes out that a hybrid subsidy policy outperforms only subsidizing consumers.

Next, for data center supply chain, this book focuses on the data center network design and resource provisioning. Data centers are the physical infrastructure for Internet related service and cloud computing. It often involves high financial inputs, which may not only affect the short-term operational level decisions but also long-term strategical level decisions. Therefore, it is crucial to construct an integrated mathematical

model，which covers electricity cost，fixed cost，latency cost and so on. The whole model is a hard-to-solve mixed integer non-linear programming problem，which motivates us to design two Lagrangian based algorithms to improve computational performances. By collecting real-word data to design a data center network for the U.S.，fruitful network design guidelines are generated after sensitivity analysis.

Finally，for cloud computing supply chain，this book focuses on resource provisioning problem under service level agreement (SLA). SLA is a common type of contract for cloud computing supply chain，where service providers need to guarantee their service quality during the contract，or provide compensate otherwise. Thanks to the development of virtual machine technology，service providers are able to monitor and adjust the resource in real time. Therefore，it is crucial to construct a robust dynamic programming model to optimize the dynamic resource adjustment problem. After providing a convexation algorithm to convert value function in each period into a piece-wise linear convex function，the robust dynamic programming problem can be solved with several linear programming problems. By generating data by using queuing network approximation，it is inspiring to find that dynamic adjustment has great potential to reduce total cost.

**Key Words:** New energy vehicle supply chain; Data center supply chain; Resource allocation; Integer programming; Robust optimization

# 目 录

# 第 1 章　引　　言

## 1.1　研 究 背 景

21 世纪以来，迅猛发展的科技改变了人们的生活方式，同时也为传统的供应链管理带来了新的挑战。一方面，新技术的不确定性对管理者的决策工具提出了更高的要求；另一方面，丰富的数据也为管理者做出合理的决策提供了有利的条件。因此，新技术下的供应链向决策制定者提出了挑战与机遇并存的运营管理新问题。

供应链管理通常涉及对整个组织网络中所有活动的管理，为最终客户提供商品或服务。供应链上活动的效率可能会对客户的满意度及供应链上的成本带来巨大影响。因此，传统的供应链管理就是要协调供应链上不同主体间的关系，通过降低库存、优化网络结构、合同制定等方法降低供应链上的运营成本。然而近些年来，从业者和学者们发现供应链管理不仅仅是降低成本的手段，而且可以成为竞争优势的来源。比如沃尔玛正是凭借其高超的供应链管理手段，一方面能够保证货品的稳定供应，另一方面还能够以低价吸引更多的消费者。

随着新兴技术的发展，一些传统供应链难以涉及的领域，比如 IT 供应链、新能源汽车供应链，为管理者带来了新的挑战（施耐德 • 劳伦斯、申作军，2016）。新的技术往往能给企业带来初期的竞争优势，然而新技术的不确定性也为运营管理带来了挑战。在新兴技术下的供应链中，由于新技术的不确定性，相比传统供应链，企业往往更加难以准确估计用户的需求，而用户对产品或者服务的要求却在逐渐提高。一些企业虽然极具创新精神，在技术方面有着先发优势，却由于运营管理不善，失去了领先地位。我国台湾宏达电子（HTC）是世界上首个制造搭载安卓系统的手

机厂商，手机市场份额也一度位居前列，比如 2012 年的全球智能手机销售市场，HTC 的销量仅次于三星、苹果、诺基亚，位列第四。然而 2019 年 5 月 10 日，HTC 在中国大陆的旗舰店已全部关闭。截至 2020 年 2 月 7 日，HTC 中国官方社区正式关闭，基本宣布退出了中国大陆市场。智能手机在 2000 年左右算是当时的新兴技术，HTC 也抓住了搭载这一技术的契机，然而现今却不得不将自己一半的设计和研发团队卖给了谷歌。其中一个很重要的原因在于其对供应链的掌控能力欠佳，比如 2010 年 8 月屏幕供应商三星由于产能不足对 HTC 断供，HTC 不得不临时更换为索尼供应。这次事件导致了最终产品的供货量难以满足需求，导致供应链遭到了巨大损失。对资源的分配失衡、资源分配策略的低效成为 HTC 手机业务走向颓势的重要原因之一。①

虽然在新兴技术的影响下，对供应链的管理充满了挑战，但是另一方面，随着大数据时代的到来、供应链中的信息系统愈加成熟，企业可以从海量数据中刻画消费者的偏好从而辅助决策，这也为供应链的管理带来了机遇。数据驱动决策，更加精细化、鲁棒性更强的运营管理就成了学界和业界的研究重点。合理地利用数据有利于设计高效率的供应链，从而降低运营成本，提高企业的竞争力，同时进一步推进新技术的普及。来自斯坦福大学的 Hau L. Lee 教授在其“The New Supply Chain Renainssance”讲座中也提到，行业头部的企业应该拥抱新技术以提高供应链效率，加强合作关系以保证供应链的集成，创新商业模式以创造更多社会价值。②

资源分配是供应链管理中非常重要的研究方向。熊彼特之后的经济学家的一个共识就是科技创新与资源分配相结合是维持经济长期增长的主要动力（Kogan et al.，2017）。微观上的低效资源分配可能导致宏观上的全要素生产率（total factor productivity）的降低，从而影响整个国家的财富水平（Hsieh & Klenow，2009; Acemoglu et al.，2018）。对于新兴技术下的供应链而言，由于其新颖性，大部分管理人员缺乏相应的管理经验，其面临的资源有效分配问题显得尤为突出。这里的资源既包括硬件、金钱等有形资源，也包括人力资源这种无形资源。本书主要聚焦于有形资

① 极客公园，网址为 https://www.geekpark.net/news/200576。

② 香港大学系统分析全球领袖研讨会，网址为 https://www.saleaders.hku.hk/hau-lee。

源的分配。

新兴技术一般包括 5G 技术、物联网技术、人工智能、可控核聚变、基因工程、云计算、便携新能源等。[①] 不同领域的新兴技术特点各不相同，其面临的资源管理问题也风格迥异。因此，本书选取具有代表性并且极具潜力的新兴技术，聚焦于新能源汽车供应链与数据中心供应链，并具体针对供应链中的资源分配问题进行研究。

新能源汽车能够有效减轻传统汽油车带来的空气污染问题，也被认为是最有潜力的能够占有未来汽车市场的出行方式。根据国际能源署的统计，2018 年的电动汽车一共消耗了 58 太瓦·时，相比使用内燃机类型的汽车缩减了 360 万吨的二氧化碳排放。[②] 中国政府一直在大力推广电动汽车的普及，在过去十年中国投入了近 2100 亿元的补贴。自 2015 年以来，纯电动汽车的产销量达到 24 万辆，2018 年的产销量超过 125 万辆，中国电动汽车一直位列全球产销量第一，而且一直保持高速增长。[③] 电动汽车与现在备受青睐的自动驾驶、汽车智能化结合得更加紧密，电动汽车很可能为传统汽车行业带来翻天覆地的变化。因此，设计高效的补贴分配策略能够有效促进新能源汽车供应链的发展。

数据中心是云计算及其他互联网相关服务的物理基础。云计算被认为是计算机领域的重大革新，云计算指的是用户无须直接管理硬件，但可以按需求直接从网络获取包括存储和计算在内的计算机资源的技术。自从亚马逊 2006 年提出弹性计算网络（Elastic Compute Cloud）以来，云计算逐渐风靡全球。云计算本身也成为其他新兴技术，如工业互联网、物联网、人工智能、区块链等技术的载体，成为这些新兴技术的主要驱动力。据《中国云计算产业发展白皮书》，2015 年的中国云计算产业规模达到 387.3 亿元，而截至 2019 年，中国云计算产业规模达到了 1290.7 亿元，年增长率一直保持在 30%左右。虽然中国云计算产业的规模一直保持较高增速，但其发展潜力还远远没有被发掘出来。2018 年的中国云计

① TechRepublic，网址为 https://www.techrepublic.com/article/top-10-emerging-technologies-of-2019/；世界经济论坛，网址为 https://www.weforum.org/agenda/2019/07/ these-are-the-top-10-emerging-technologies-of-2019/；MIT 科技评论，网址为 https://www.technologyreview.com/lists/technologies/2019/。

② IEA 全球电动车市场调研，网址为 https://www.iea.org/reports/global-ev-outlook-2019。

③ 前瞻产业研究院，网址为 https://bg.qianzhan.com/trends/detail/506/190429-15886178.html。

算市场规模仅达到了美国市场的 8%左右，仍有很大发展空间。[①] 数据中心是各种互联网服务的物理基础，无论是电子商务、搜索、流媒体服务，还是云计算，背后都需要完善稳健的数据中心支持。数据中心的能力也被作为衡量企业未来发展潜力的重要指标之一。因此，优秀的数据中心网络设计及合理的服务资源供给策略能够有效提高数据中心供应链的运作效率，提高其行业竞争力。

供应链管理包括企业不同层级的决策，既涉及具有长期影响力的战略层级，也涉及每年或每季度更新的战术层级，同时也涵盖了每天甚至适时调整的运营层级。Simchi-Levi et al.（2008）总结了供应链中常见的问题，其中包括分配中心的网络设计、库存控制、供应链机制设计、产品设计等。因此本书涉及的数据中心网络设计、电动汽车的补贴资源分配，以及云计算服务器的资源供给都是供应链管理所涉及的重要问题。然而由于数据中心供应链与新能源汽车供应链涉及如云计算和电动汽车这样的新兴技术，其供应链管理也具有不同的管理难点。比如：新能源汽车供应链中需要考虑电动汽车与充电桩之间的网络效应，而传统汽车的加油站相对比较密集，因此传统汽车供应链中很少考虑两者间的网络效应；数据中心供应链中的网络设计与服务资源分配需要考虑数据处理的延迟，而传统的分配中心选址往往会忽略与之对应的配送中心的内部处理时间；数据中心供应链中，可以动态调整云计算服务资源来满足服务水平协议的要求，服务资源的运行状态与传统供应链的先预测、再做决策的方法更有显著差异。这些新兴技术的供应链，一方面与传统供应链的特点风格迥异，另一方面其管理人员缺乏相应的管理经验，因此使用合理的分析建模工具进行有效的资源分配显得尤为重要。

综上所述，本书将聚焦于新能源汽车供应链与数据中心供应链，并对相关运营管理中的资源分配问题进行研究，本书将回答以下研究问题：

- 新能源汽车供应链中电动汽车的补贴资源分配问题。电动汽车作为新兴技术，其初期发展依赖于政府的补贴。消费者对电动汽车的偏好不仅取决于政府对电动汽车消费者补贴的力度，同时也受到充电桩数量的影响。因此本部分构建一个分析模型，综合考虑电动汽车消费者、充电桩投资者及政府的效用，优化政府的补贴

① 《中国云计算产业发展白皮书》，网址为 https://www.sciping.com/33520.html。

资源分配问题。最后，本部分通过实际数据比较了单补贴政策与双补贴政策的优劣。

- 数据中心供应链中的网络设计与服务资源分配问题。数据中心网络是互联网相关服务的物理基础。合理的数据中心网络设计能够有效降低运营方的总成本，同时可以提高使用服务的消费者的满意度。本部分针对数据中心网络设计与基础设施资源分配问题构建了数学规划模型，通过优化数据中心的位置和服务资源分配决策，来降低总运营成本和服务延迟损失。最后，通过实际数值实验的结果来指导实践中的数据中心网络设计。
- 数据中心供应链中的云计算备用服务资源供给问题。对于云计算服务行业，一般更加注重需求管理，服务提供商一般需要与用户签订服务水平协议来保证提供持续稳定的服务，如果没有达到协定服务水平则需要付出违约成本。因此本部分考虑了云计算服务器的运行不确定性，使用数据驱动的方法构建鲁棒动态优化模型，通过动态调整服务资源，最小化服务资源占用成本，同时保证协议中的服务水平。最后，通过生成云计算服务器宕机时间历史数据，应用本模型证实动态调整服务资源的优势。

图 1-1 阐释了本书研究的内在联系。本书聚焦以新能源汽车和数据中心为例的供应链中的资源分配问题。从研究话题上来说，本书三部分都围绕着新兴技术下的供应链中的资源分配问题展开，具体研究内容包括政府财政补贴资源的分配、数据中心计算存储资源的分配，以及云计算服务提供商备用服务器资源的分配。三部分涉及的主要研究内容都与“新基建”计划的基础设施建设密切相关，具有重要的现实意义。[①]“新基建”计划是与传统的“铁公基”相对应的，是夯实新一轮的科技革命高质量发展的基础。与传统基建相比，“新基建”更要突出支撑产业升级、政府的软治理及生产要素的整合和分配（赛迪智库电子信息研究所，2020）。由于“新基建”具有新颖性，很多管理人员缺乏相关的经验，因此这也为相关管理人员带来了更大的挑战。

从研究方法上来说，本书研究的三个子问题都结合了现实数据，由理论联系实际。问题涉及的数据量级随着决策层级从战略层面到具体的运

① 新华网，网址为 http://www.xinhuanet.com/politics/2020-04/26/c_1125908061.htm。

营层面逐步增大，既包括了传统的模型驱动决策（第 3 章和第 4 章），也包含了较为新颖的数据驱动决策（第 5 章）。为了解决具体的研究问题，本书综合使用了管科领域中的不同工具，例如博弈论、混合整数规划、鲁棒动态优化等，量体裁衣，能够为实际问题提供更加完善的解决方案。

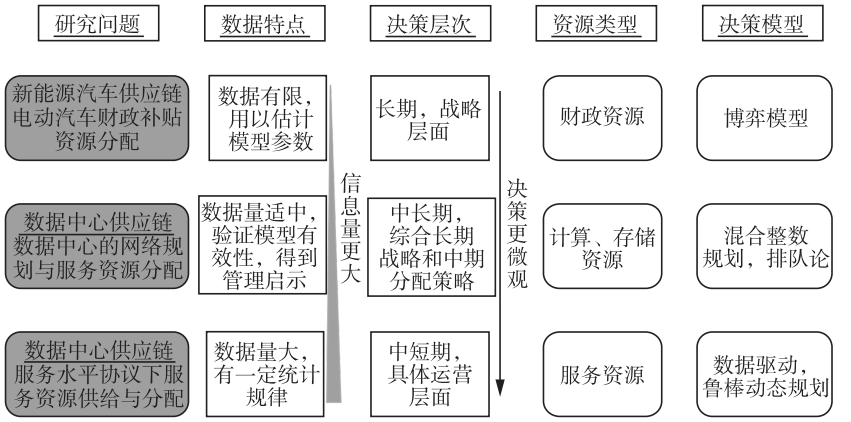



**图 1-1 研究问题的内在联系**

## 1.2 研 究 意 义

新兴技术的新颖性及管理人员缺乏相关的经验，为新能源汽车供应链带来了更多不确定性。本书聚焦的三个方面——新能源汽车供应链中电动汽车补贴分配政策、数据中心供应链中网络设计及服务水平协议下的云计算备用服务器资源管理，在目前都存在优化改进的空间。

电动汽车由于其低噪声、低污染、技术相对成熟稳定等特点受到很多国家政府的青睐。多国政府都出台了补贴、税费减免等优惠政策来促进电动汽车的推广。然而目前电动汽车的市场占有率尚不尽如人意。根据国际能源署的统计，目前挪威的电动汽车市场占有率为世界第一，约为 46%，而第二名冰岛和第三名瑞典的电动汽车市场占有率分别只有 17%和 8%。[①] 虽然电动汽车行业发展迅猛，技术日益成熟，但是目前消费者对于电动汽车仍心存疑虑，比如担心电动汽车行驶里程不足、路途中没有公用充电桩及时充电带来的里程焦虑（range anxiety），因此电动汽车的推广目前

① https://www.iea.org/reports/global-ev-outlook-2019

仍然主要依赖政策的扶持。政府既可以选择直接补贴消费者，也可以通过补贴充电桩投资者以间接缓解消费者的里程焦虑从而获得更高的推广水平。一般来说，消费者和充电桩之间有着相互的网络效应，更多的充电桩会降低消费者的里程焦虑，提高消费者的效用；更多的消费者可以增加充电桩的利润，提高投资者的效用。两者形成了正反馈。但是在市场尚不成熟的初期阶段，无论是消费者数量还是充电桩的数量都不足，形成的正反馈不够强，如果不进行有效的补贴，市场可能无法有效增长，反而会萎缩，甚至消亡。在 2010 年，德国政府定下了在 2020 年电动汽车保有量达到 100 万辆的目标，并在 2016 年启动了一个价值 10 亿欧元的电动汽车补贴政策。然而截至 2019 年 10 月，德国电动汽车的保有量仅仅达到了 142805 辆，与目标相距甚远。[①] 因此，只有合理设计电动汽车的财政补贴资源分配策略，才能达到事半功倍的效果。

数据中心是各种互联网相关服务的物理基础，其中放置了计算机及相关辅助设备，包括服务器、电源、电信设备、散热系统、安全设施等。数据中心建造成本和维护成本极高，据估计每平方英尺的建造价格高达 200∼1000 美元。[②] 尽管价格高昂，但互联网巨头们都在抢夺市场，无论是美国的谷歌、微软、亚马逊，还是国内的阿里、腾讯、百度等都在自建数据中心网络。数据中心网络最重要的意义就是为顾客提供稳定快速的服务，然而目前一些数据中心的表现仍然不尽如人意。比如苹果公司提供的 iCloud 云储存服务就一直广受用户诟病，一直到 2018 年 3 月，中国区 iCloud 正式迁移到云上贵州，这一情况才有所改善，但即使如此，糟糕的用户体验也为苹果公司带来了间接的客户流失和经济损失。2020 年新型冠状病毒疫情期间，很多线上教育平台兴起。这类直播服务一般对稳定性的要求很高，然而类似雨课堂之类的新平台由于缺乏处理高并发需求和管理服务器的经验，在运营初期屡次发生服务器错误、突然直播中断等状况，为新产品的推广带来了严重的负面影响。因此，设计高效的数据中心网络，成了每一个提供云服务企业所面临的问题。数据中心网络设计包括数据中心的选址、数据中心资源对需求点的分配、数据中心

① 数据来源：https://www.kba.de。

② https://www.theengineeringprojects.com/2018/12/what-are-data-centers-and-how-much-do-they-cost.html

内部资源分配等。对资源分配的优化有利于降低数据中心网络的总成本，并且也能在控制成本的同时提供高质量的服务。

在数据中心供应链中，服务水平协议是云计算服务中常用的一种合作方式。随着中国的经济发展，传统的需求管理已经很难跟上市场的要求，供给侧需要有效提高其服务质量来支撑其市场规模。一般来说，客户与服务供给方之间签订的服务水平协议会定义服务的内容及双方的权利和义务。常见的服务水平协议中会规定供给方服务的标准、服务的可用性、服务质量等。在数据中心供应链的云计算场景中，服务水平协议尤为普遍。服务提供商通过将数据中心基础设施虚拟化，使用网络来服务消费者，其重点在于提供持续稳定的服务。据估计，世界排名前三名的云服务提供商如果服务器停摆三天有可能带来 47 亿 ～ 69 亿美元的损失。[①] 诚然，这样大型的云计算服务提供商很少会出现服务器宕机的状况，然而黑客攻击、自然灾害、硬件故障等难以避免的状况常常会导致云计算服务的中断。随着云计算服务的发展，顾客对云计算的需求就如同对水、电、网络的需求一样，即使短时间的服务中断也会给顾客带来很强的不信任感。云计算服务的宕机可能对企业造成非常严重的影响，所以一般都会提供一些备用服务器以方便随时切换和恢复服务。在服务水平协议下，合理分配备用服务器资源能够有效平衡服务器费用和宕机带来的损失，达到成本效益最大化。

综上所述，新能源汽车供应链与数据中心供应链影响着人们生活的方方面面，同时也为其涉及的资源分配管理问题带来了很多挑战。对资源分配策略的优化能够在保证服务质量的同时，有效降低运营成本，因此本书试图通过优化有形资源的分配，解决供应链中存在的运营管理问题。

## 1.3 研 究 框 架

本研究聚焦于供应链中的资源分配等运营管理问题。不同供应链由于其涉及的具体技术不同，其管理难点也各不相同。因此本书选取近些年来比较有代表性的，并可能给未来社会带来变革的数据中心供应链与新

① https://www.zdnet.com/article/cloud-computing-heres-how-much-a-huge-outage-could-cost-you/

能源汽车供应链，聚焦于电动汽车市场和云计算市场所面临的有形资源的分配问题。第一部分的新能源汽车供应链的财政补贴政策，主要阐述政府应如何分配财政资源以补贴消费者或者充电桩投资者；第二部分的数据中心网络布局和资源分配，主要阐述数据中心如何向消费者分配其计算、存储等服务器资源；第三部分的服务水平协议下的云计算备用服务资源供给，主要阐述云计算服务提供商如何分配备用服务器资源给需求方。图 1-2 总结了本书的研究框架，接下来将针对具体问题详细介绍每部分的研究内容。

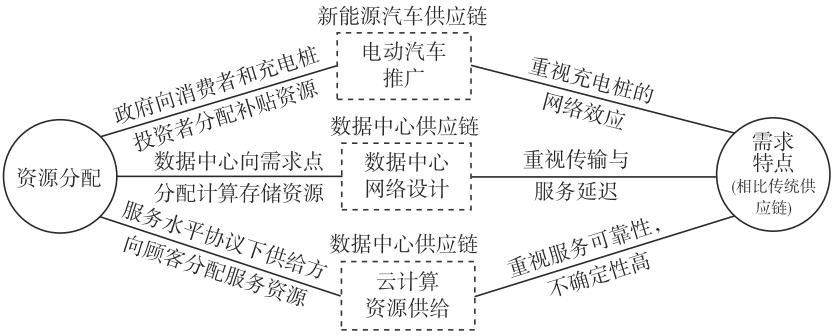



**图 1-2　研究框架**

对于新能源汽车供应链的电动汽车的补贴政策问题，政府需要合理分配财政补贴资源，最大化提高电动汽车的市场占有率。政府的补贴既可以直接给消费者，使得消费者实际购买电动汽车的价格降低；也可以通过补贴充电桩投资者，使得市场上最终建成更多的充电桩，缓解消费者的里程焦虑，间接提高消费者使用电动汽车的效用。因此，在政府进行决策的时候，需要同时考虑消费者和充电桩投资者的效用。对于消费者而言，电动汽车带来的效用主要由电动汽车价格、补贴金额、充电保养费用、充电桩带来的网络效应，以及消费者个人对电动汽车的偏好构成。对于充电桩投资者而言，需要权衡补贴后的建造充电桩的成本及可能由消费者带来的利润，以决定是否要建造充电桩。综合考虑消费者和充电桩投资者的效用后，可以构建出政府所面临的优化问题，权衡电动汽车给政府带来的好处和补贴带来的成本。接下来本部分对模型进行了求解，通过分析解的性质，可以得到一些有意义并且一定程度上反直觉的结论。比如当充电桩的建造成本更高的时候，政府应该将更多的补贴放到消费者端，而不是投资

者端。在消费者的环保意识更强的时候，政府可能减少对消费者和充电桩的补贴，最后导致电动汽车数量反而降低，等等。通过分析这些性质背后的原理，能给政策制定者以启示。由于一些政府只对消费者进行补贴，本部分也对这种补贴政策进行了研究，并与提供双补贴政策的模型进行了对比。结果发现，使用双补贴政策总比使用单补贴政策会达到更好的效果（也可能相同），尤其当充电桩的建造成本不高的时候，其优势更加明显。本部分还搜集了深圳的电动汽车市场相关数据对模型进行了校正。代入实际数据发现，使用双补贴政策的效果要远优于使用单补贴政策，此部分内容将在第 3 章中详细阐述。

对于数据中心供应链的网络布局及资源分配问题，不同于传统供应链的仓库选址问题，在考虑数据中心的固定成本之外，还需要考虑数据中心的其他特点带来的成本。首先，数据中心带来了不容忽视的能量消耗，据估计，当今的数据中心消耗了全球 2%的电量，并且预计在 2030 年可能达到全球总电量的 8%，[①] 因此在目标函数中还需要考虑数据中心的耗电成本。其次，数据中心的服务质量也是影响服务提供商决策的重要问题，本部分考虑了数据中心的延迟成本，其中既包括数据中心到需求点之间的传输延迟，也包括在数据中心内处理的终端延迟。比如苹果公司之前的服务器放置在美国本土，中国用户访问软件商店或者其他存储服务的时候就会产生传输延迟；而类似地，淘宝网在刚刚创立“双十一”之际，由于数据中心的服务器计算能力不足，很多人下订单时遇到延迟甚至网站崩溃的情况。这两种延迟都会降低用户的满意度及未来需求，影响服务提供者的商誉，因此在目标函数中也应该考虑这两种延迟成本。再次，很多时候服务提供商还会使用托管服务中心。为了提供更好的服务，有时服务提供商会向其他数据中心建造者租赁一部分服务器用于服务当地的用户。对于这部分托管数据中心，同样也需要考虑其固定成本（租赁成本）、耗电成本及终端延迟成本。由于通常托管数据中心仅服务当地的需求，其传输延迟一般可以忽略不计。最后，除基础模型外，本部分还考虑了一些模型的拓展，例如考虑需求点之间可能有相互依赖性、电能消耗与使用率的非线性关系、大量需求带来的网络拥堵等。数据中心的建造者可以通

① https://fortune.com/2019/09/18/internet-cloud-server-data-center-energy-consumption-renewable-coal/

过调整数据中心的地理位置，数据中心内部的存储、计算等资源的配置，以及对具体需求点提供的资源量，优化上述考虑的总成本。综合考虑数据中心网络所涉及的成本及其面临的约束，可以将该优化问题转化为一个非线性整数规划模型。这类模型一般都是 NP 难问题，计算复杂度比较高。然而，无论是基础模型还是拓展模型，都可以将其转化为整数二阶锥优化问题，可以直接使用商业软件求解。但是如果问题的规模比较大，直接使用商业软件求解，计算时间仍然很长，甚至无法得到可行解。因此，本书还针对问题的结构提出了一个适用于所有模型的拉格朗日算法，以及一个适用于基础模型的加入加强割的拉格朗日算法。在构建完模型并针对问题结构设计了算法后，本部分进行了数值实验。一方面，采集了美国当前的电价、数据中心固定成本等数据，并利用公开数据估计了不同地区的需求率，使用实际数据设计了适用于美国本土的数据中心网络，并据此提出了一些构建数据中心网络的建议；另一方面，本部分也随机生成了不同规模的数据中心网络数据，分别直接使用 Gurobi、使用两种提出的拉格朗日算法对算例进行计算，并对计算效果进行了比较。结果表明，提出的两种算法要比直接使用 Gurobi 的计算效率高得多，尤其是问题的规模比较大、电量约束比较紧的时候，使用提出的拉格朗日算法效果更好，此部分内容将在第 4 章详细阐述。

对于数据中心供应链中的云计算备用服务器资源供给策略问题，云计算服务提供商需要考虑需求和服务器运行的不确定性，动态调整备用服务资源的分配策略。一般来说，在数据中心供应链中，云计算服务提供商在提供服务前会与顾客签订一个服务水平协议，其中要求在服务提供商合同期间绝大部分时间可以满足顾客的需求，否则就要对顾客进行赔偿（Yuan et al.，2018）。如今先进的物联网和互联网技术使得服务供给方可以实时观测服务器的运行状态并迅速调整资源供给水平。因此服务供给方可以通过记录的累计宕机时间，动态调整其服务资源供给策略。这个问题可以被构造成一个动态规划问题，但与传统动态规划不同的是，服务资源运行具有不确定性，并且难以直接估计分布。因此本书采用了数据驱动的鲁棒优化方法，通过历史数据，使用 Wasserstein 距离来刻画参数的不确定集。这样一来，动态调整服务资源的问题可以被构造成一个鲁棒动态规划问题。为了求解此问题，本书使用逆向归纳的方法，首先解决最后

一期的问题。最后一期的鲁棒优化问题可以转化为有限维的线性规划问题进行求解，但是最后一期的价值函数关于状态一般是一个非凸的函数，代入倒数第二期的问题中难以直接求解。因此本书构造了一个凸化算法，将价值函数转化为一个分段线性的凸增函数，这样一来，倒数第二期的优化问题也可以转化为有限维的线性规划问题进行求解。对每一期的问题都采用此凸化算法，就可以使用线性规划的方法解决前一期的优化问题。对于这个有限期的鲁棒优化问题，可以通过解决若干个线性规划的问题得到满意的资源供给策略。最后，本书还将提出的算法应用于云计算的备用服务器供给问题。通过案例研究可以发现，鲁棒动态规划模型求解出来最优策略要比传统的静态策略降低将近 1/4 的总成本，此部分内容将在第 5 章中详细阐述。

第 6 章总结了本书的主要工作与贡献。主要聚焦于新能源汽车供应链与数据中心供应链，并对其面临的有形资源分配问题进行了建模求解。三个部分的研究问题涉及了不同的研究对象和不同的数据量，因此使用了不同的理论方法，通过数值计算或者理论推导找到一些模型解的性质，以此对现实供应链的管理带来启示。不仅如此，在求解模型的过程中，也提出了一些新颖的算法，填补了过去文献中的理论空缺。

# 第 2 章　文 献 综 述

本书聚焦于供应链中的资源分配等运营管理问题。本章将首先对供应链中的资源分配相关文献进行梳理，然后具体介绍每部分的具体情景管理问题，以及建模求解过程中所使用理论工具的相关文献。

供应链中面临着资源分配问题，不合理的资源分配可能导致新兴技术的转化效率降低，从而影响其发展。在社会动力学中，有一个关键节点（critical mass）问题，指的就是社会系统中的创新需要一定比例的接受才能够保证该创新自我维持并增长。只有合理的分配资源才能保证新兴技术的不断发展，否则即使该技术有重大意义，也可能中途夭折。比如谷歌公司就已经舍弃将近 200 项曾经投资巨大的创新项目。[①] Hsieh & Klenow（2009）利用微观企业数据分析对比了中国、印度与美国的资源分配不均的状况，发现资源的不合理分配可能会降低社会全要素生产率。Acemoglu et al.（2018）提出了一个预测企业行为的模型，将资源再分配与企业创新结合起来，使用了美国普查局的数据检验了企业创新、生产力增长及再分配的影响因素。

接下来，本书将分章节具体介绍每部分研究问题的相关参考文献。其中对于新能源汽车供应链中的财政补贴资源分配问题，将介绍消费者对电动汽车偏好的相关研究、电动汽车补贴政策对电动汽车市场的影响及一些优化补贴政策的相关研究；关于数据中心供应链的网络设计与资源分配问题，将介绍选址问题的研究历史及本部分所使用的二阶锥优化问题的研究进展；对于服务水平协议的云计算服务资源供给问题，主要介绍本书所使用的鲁棒动态规划相关研究。图 2-1 总结了本部分文献综述的结构。

① https://killedbygoogle.com/

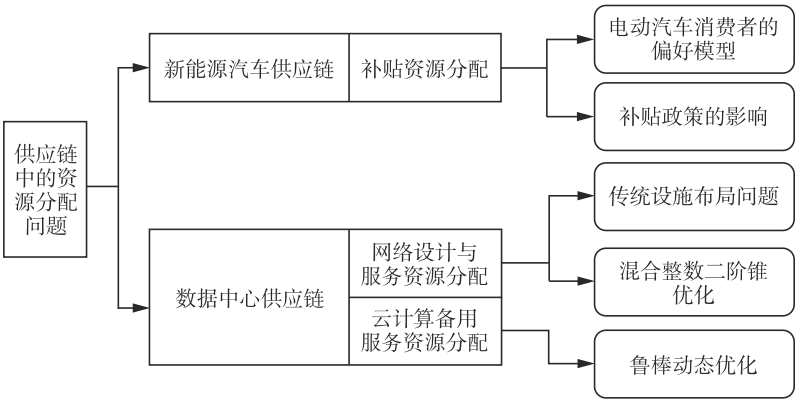



**图 2-1 文献综述结构**

## 2.1 新能源汽车供应链补贴资源分配相关文献

电动汽车属于比较新的技术，很多消费者对电动汽车仍持怀疑态度，因此一部分文献采用实证等方法找到影响消费者购买电动汽车的因素。Helveston et al.（2015）通过问卷调查，对比了中国和美国的消费者对电动汽车的偏好，结果发现中国消费者偏好中距离的插入式电动汽车，而美国消费者即便在有补贴的情况下仍然偏好短距离的插入式电动汽车。文章还发现中国采用电动汽车可以降低对汽油的消耗，并且减少温室气体排放，而中国电动汽车市场的蓬勃发展也吸引世界其他发电碳排放较低的国家更多采用电动汽车。Han et al.（2017）使用消费价值理论将消费者对电动汽车的偏好分为了包括价钱、性能和便捷性的功能性价值，以及包括情感、社会和认知的非功能性价值。通过对中国合肥 607 个司机进行问卷调查得知，功能性价值对顾客的购买行为有着直接和间接的影响，但是非功能性价值只能通过态度对购买行为有间接影响。White & Sintov（2017）检验了电动汽车可能体现的驾驶者的自我身份认同。消费者的环保精神成了预测消费者是否购买电动汽车的最强的指标，而消费者的好奇心能够更好地预测是否租借或者购买电动汽车，这两个指标要比其他人口统计指标的影响强很多。Degirmenci & Breitner（2017）采访了 40 个电动汽车驾驶者并进行了 167 次邀请试驾的实验，在进行调查并使用结构方程模型估计后，发现消费者对电动汽车的环保性能评价是影响其

购买意愿的最重要的因素。Lin & Wu（2018）通过向中国几个一线城市，包括北京、上海、广州、深圳，发放调查问卷分析消费者购买电动汽车的影响因素，结果发现网络效应、电动汽车价格、政府补贴、车辆性能、环保等因素，以及性别、年龄、婚否都对消费者购买电动汽车的意愿有显著影响。

电动汽车的市场占有率与政府的政策息息相关，因此一部分文献采用实证的方法研究政府政策对电动汽车推广的影响。Mueller & Haan（2009）使用基于 agent 的仿真方法，考虑消费者进行两阶段的决策，第一阶段仅考虑电动汽车本身的特点，第二阶段考虑各种政府政策。通过仿真可以得到给定政策下的消费者购买行为及政策最后的效果。Gallagher & Muehlegger（2011）使用美国市场 2000—2006 年的数据，分析了汽油价格、消费税、收入税及消费补贴等政策对混合电动汽车的影响。Sioshansi（2012）研究了不同电价定价策略下的插入式电动汽车消费者的充电行为，并在系统运营者控制充电行为的最优方案下比较其成本和碳排放。结果发现，使用实时定价策略表现不佳，控制夜间充电比控制白天充电效果更加显著。Zhou et al.（2013）从能源消耗和温室气体排放的角度研究了中国电动汽车的发展状况，指出未来更多的电动汽车带来的能源节约和温室气体减排效果会更加明显，并为具体的电动汽车发展提出了政策建议。Jenn et al.（2013）估计了美国联邦政府自 2004 年提供的混合电动汽车相关政策的影响，结果发现 2004 年的减税未对电动汽车的销量带来明显提升，但是 2005 年的能源相关政策使不同型号的电动汽车有了 3%∼20%的销量提升。Hao et al.（2014）总结了中国的两阶段电动汽车补贴政策，并估计了该政策对电动汽车市场的影响。其文章得出结论，在 2015 年中国政府对电动汽车的补贴非常必要，而且当前的补贴上限不足。在 2015—2020 年，电动汽车的发展可能减少对补贴的依赖。Gnann et al.（2015）研究并预测了德国的插入式电动汽车的市场演变，结果发现能源价格对电动汽车存量影响较大，如果汽油价格上涨 25%，可能会吸引双倍的人购买插入式电动汽车。其文章认为，由于市场演化具有很强的不确定性，因此政策应该随之动态调整。Holland et al.（2016）结合了消费者购买电动汽车的模型、电力导致排放的模型和一个空气污染模型来估计电动汽车对不同区域带来的环境影响。结果发现不同区域采取不同的

补贴政策可能会减少无谓损失。Li et al.（2017）考虑了充电桩与消费者之间的非直接网络效应，并使用美国 2011—2013 年的数据实证模型验证了网络效应的显著性。结果表明，如果美国政府在相同的财政支出下转而支持充电桩的建设，结果可能会比当前的推广效果好两倍。Liu et al.（2017）试图通过构建一个进化博弈的模型解释为何中国政府在诉诸诸多电动汽车补贴政策后依然没有达到预期目标。仿真结果表明，综合使用动态的排放税收政策和静态的补贴政策能更有效地促进电动汽车行业发展。Du et al.（2017）通过分析电动汽车技术，使用市场分析、专家采访等方法，总结中国电动汽车市场的优缺点。文章发现，对于电动汽车发展最重要的是电池能量密度和电池的使用寿命。中国政府应该加大力度发展安全可循环的电动汽车技术。Zhang & Bai（2017）总结了中国 2006—2016 年的新能源汽车补贴政策，并使用了政策依赖映射法分析了 175 个全国和地区性的新能源政策的影响，探讨了不同地区的发展差异及骗补现象，有助于管理者设计更好的促进措施。Zhang et al.（2017）通过搜集和分类中国的新能源汽车相关政策来估计这些政策的绩效。文章将这些措施分为三类，分别是经济措施、基础设施建设、研究和发展（R&D）投资。结果发现目前的措施仍然有待提高，需要建立充电桩和统一电价的标准等。Wang et al.（2017a）使用 2013—2014 年的 41 个城市的数据，拟合多重线性回归来分析电动汽车相关措施对电动汽车销量的影响，结果发现充电桩数量、牌照费减免、无驾驶限制及对充电桩建设提供支持是最重要的四个影响因素。Wang et al.（2017b）使用离散选择实验调查了 247 个被试并使用了混合逻辑回归来估计几种可能的电动车政策的有效性。结果发现购买限制和行驶限制是最有效的两个政策，而充电费用减免也有很好的效果，但是类似于减少停车费这类措施的有效性不强。Wang et al.（2017c）将电动车相关政策分为了财政相关政策、信息供给相关政策和便捷性相关政策，通过 324 份调查问卷，发现便捷性相关政策的影响最大，而且消费者的环保理念对电动汽车的普及也有着很强的影响。

除了实证类型文章之外，近些年也有越来越多的学者使用理论分析模型对电动汽车的补贴政策进行研究。Huang et al.（2013）考虑了一个传统燃料汽车供应链和电动汽车供应链双寡头竞争的环境，其中政府对电动汽车的消费者进行补贴。结果发现，当消费者的议价水平较高的时

候，政府的补贴策略更加有效。如果充电桩的数量足够多的话，可以减轻补贴对传统汽车供应链的冲击。Luo et al. （2014b）研究了一个生产商和零售电商的电动汽车供应链，政府为消费者提供了有上限的价格折扣的补贴政策。结果发现价格上限对单位生产成本较高的生产商影响更大，而折扣力度对单位生产成本较低的生产商影响更大。而且期望销量随着折扣力度增加而上升，但是有可能随着价格上限的增长而降低。Shao et al.（2017）分析了寡头和双寡头市场下的电动汽车和汽油车的市场，并比较了消费补贴和价格折扣两种政策的影响，结果发现在消费者剩余、环境影响及社会福利相同的情况下，政府更愿意采用消费补贴，因为这样带来了更低的支出。Springel（2016）使用挪威的电动汽车市场数据，采用了双边市场框架，考虑了消费者和充电桩之间的网络效应，构建了回归模型。结果发现 2010—2015 年对充电桩补贴对电动汽车购买带来的提升是对消费者直接补贴的 2.16 倍，然而随着对充电桩补贴的增加，这种现象会逐渐减弱。Chemama et al.（2019）研究了政府与例如电动汽车等绿色技术的制造商，在面临不确定需求时的两阶段博弈模型。结果表明，如果政府采用一个固定的补贴政策会鼓励制造商在初期生产更多。如果采用更灵活的补贴政策，制造商的预期收益会有所提高，而消费者根据需求弹性收益也有可能受损。Ma et al.（2019）考虑了类似电动汽车这类清洁能源在推广初期面临的“鸡生蛋，蛋生鸡”的问题，文章发现最优的补贴政策应当是充电桩的成本非常高或者非常低的时候仅提供对消费者的补贴，而当充电桩的成本适中时，应该既补贴充电桩投资者，也补贴消费者。

## 2.2 数据中心供应链网络布局与资源分配相关文献

数据中心的网络设计与传统的供应链选址问题有着密不可分的关系，因此本部分首先梳理了传统供应链选址问题的相关文献。历史上首个对选址问题进行研究的是 Launhardt & Bewley（1900），研究了一个工厂在三个煤矿间运煤，应该如何确定最优工厂位置，该文使用了平面几何的方法对问题进行求解。Weber（1929）也处理了类似的问题，使用了不同的方法但是得到了相同的结论。不仅如此，Weber（1929）对问题进行了拓展，考虑了超过三个煤矿的情况下的工厂最优的选址问题。对于 Weber 问题

的一个比较自然的拓展，则是同时考虑了多个设施的选址问题。Cooper（1963）引入了选址问题中的经典问题——p-median 问题，也就是每个需求点都需要被 p 个设施中的一个所服务。Hakimi（1964、1965）使用了图的绝对中心（absolute median）概念，解决了 p-median 问题。该文使用的性质意味着很多选址问题都可以转换成离散的设定，从而可以使用整数优化和组合优化的技巧求解这类问题。Balinski （1965）首次将选址问题构造成了一个混合整数线性规划问题（MILP），该问题则是经典的无容量限制的设施选址问题（UFLP）。这些早期的文章构筑了选址问题的基础，支撑着近五十年来该领域的蓬勃发展（Laporte et al.，2015）。

数据中心与传统的仓储中心不同，仓储中心一般是由仓储中心出发，向需求点提供服务，而与之相反的，数据中心属于不可移动的服务端，一般是数据传输到数据中心内进行处理。因此仓储中心这种设施选址，一般不会考虑仓储中心内部的等待延迟，但是数据中心这种不可移动的设施，则需要考虑对应的终端延迟成本。Berman & Krass（2015）对这类需求随机、考虑拥堵而且服务端不可移动的模型（SLCIS）进行了全面而细致的总结。SLCIS 类型的模型考虑了包括顾客、设施及其交互关系的系统。具体来说，对于顾客，一般假设顾客的集合是离散的而且数量有限，同时需求是同质的，也就是说到了设施后将不对需求进行区分；对于设施，同样也一般假设设施的集合是有限且离散的。设施为顾客提供服务，由于服务设施空间等的限制，服务率有上限，顾客在设施内形成了一个可能有多服务台的队列。每个设施都面临着需求分配的决策，也就是该服务设施需要服务哪些顾客。对于 SLCIS 模型，一般要求每个顾客都至多有一个设施对其提供服务。与可移动的服务设施相比，不可移动的设施的一个重要特点是每个设施都可以看作独立的排队系统，其原因在于对于可移动的设施来说，需求分配往往是根据系统状态动态分配的，这就导致了不同设施之间的队列存在相互依赖性，从而不能直接分割。

很多基于不同假设、约束的模型都可以认为是 SLCIS 类型的模型，虽然特点迥异，但是一般可以将其分为四类模型，分别是覆盖顾客导向的（Coverage-Oriented）、服务质量导向的（Service-Objective）、均衡目标的（Balanced-Objective）和显式顾客反馈的（Explicit Customer Response）。覆盖顾客导向的模型旨在建立一个系统为顾客提供充分（adequacy）的

服务，所谓充分一般通过距离和延迟来定义，并由覆盖和服务水平约束所控制。这类模型一般假设设施不可能同时满足所有顾客的需求，因此目标函数为最大化满足顾客的需求。服务质量导向的模型旨在使用有限资源优化客户服务，其目标函数一般考虑延迟成本和服务台资源带来的成本。如果设施的容量是可以变化的，此类模型的难度骤升，因此大部分文献考虑的设施都具有固定的容量。均衡目标的模型旨在设计一个能够平衡顾客和设施成本的系统，其中顾客需要支付旅行和等待成本，而设施建设者需要承受与设施相关的成本。显式顾客反馈的模型指的是着重刻画顾客行为的一类模型，顾客的需求具有弹性，由于等待时间或者旅行距离，顾客可能会调整其需求量。一般来说，此类模型的最优解往往是能够最大化顾客效用的解。

数据中心网络设计问题属于均衡目标类型的模型，此类模型在近些年受到了学界的广泛关注。Wang et al.（2004）受银行系统的通信网络和 ATM 机器系统的启发，考虑顾客的随机需求和拥堵，研究应如何设计固定服务设施的选址。目标函数中考虑了服务提供商的安装和操作服务器的成本、顾客等待服务的成本，该文对模型进行了变形转化，使之成为传统的设施选址问题，并提出了启发式算法。Elhedhli（2006b）优化了选址决策，将服务能力分配到具有随机需求的顾客中，该文提出了一种分段线性的近似方法和一个基于割平面的精确方法，可以解决多达 100 个顾客、20 个候选设施规模的问题。Aboolian et al.（2008）考虑了拥挤网络中的设施选址和服务器分配的问题，目标函数中包括了设施的建造固定成本、可变的服务台成本、顾客的旅行时间成本和在设施中的等待成本。该文既提出了精确算法，也提供了复杂度更低的近似算法。Castillo et al.（2009）同时考虑了顾客的旅行和延迟成本，也考虑了设施相关成本，同时解决了选址问题和需求分配问题。模型可以适用于呼叫中心、诊所、汽车检测中心等系统。该文使用了拉格朗日松弛算法，并通过算例证明其有效性。Zhang et al.（2009）在医疗设施网络中考虑了服务的拥堵情况，通过优化选址决策最大化提升对患者的服务水平。其中顾客等待时间使用了 $M/M/1$ 模型进行刻画。Zhang et al.（2010）在上述模型的基础上进行了拓展，考虑了资源容量的决策。他们使用 $M/M/s$ 排队模型来估计顾客的等待时间，并将服务台数量 $s$ 作为决策变量。Abouee-Mehrizi et

al.（2011）考虑了类似的问题，顾客按照泊松过程到达，到达率取决于价格和距离，顾客观察所有的服务设施并按照多元罗吉特模型选择其中一个设施完成服务。该文设计了一个近似算法并通过算例验证了算法有效性。Paraskevopoulos et al.（2016）研究了固定价格多商品的网络设计问题，但是在传统的运输和建造成本之外，还考虑了设施上的拥挤带来的延迟成本。该文将此问题建模为非线性整数规划问题，并将其转化为二阶锥优化问题进行求解。同时也提出了基于本地搜索和网格搜索的进化算法，并给出了算法的上界。

数据中心网络设计综合考虑了战略层面的选址类型的决策，也考虑了运营层面服务器资源供给、需求分配的决策。类似的整合模型在文献中并不鲜见。Daskin et al.（2002）综合考虑了配送中心的选址问题及其中的库存控制问题，采用了拉格朗日松弛算法，并提出了一些启发式算法解决拉格朗日子问题。最后进行了数值实验和敏感性分析验证了算法的有效性。Shen et al.（2003）考虑了与上述问题类似的问题，但是将问题转化成了一个集合覆盖问题，使用了列生成算法来解决此问题，最后也通过算例证实了该算法的有效性。Geunes & Pardalos（2003）综合考虑了供应链管理与金融工程的模型，将此类模型分为两类，第一类考虑了整合模型带来的风险分摊（risk pooling）的现象，还有一些文献考虑了金融理论的条件风险价值（conditional value at risk，cVaR）。Taaffe et al.（2008）综合考虑了营销力度、市场选择及采购决策来最优化企业的总利润，将该问题构建成混合整数规划模型，并提出了一个新颖且高效的分枝定界算法。Geunes et al.（2011）在传统供应链网络规划问题的基础上考虑了不同市场的选择。在第一阶段，决策者决定一个细分市场，在第二阶段再进行一个满足所有细分市场需求的最小化生产成本的决策。该文基于线性规划提出了一个近似算法，并证明了算法的理论界。

数据中心还有一些不同于传统供应链选址问题的特点。近些年，在管理科学和计算机领域都对此类问题进行了一些探索。Greenberg et al.（2008）认为云服务的基础建设投资消耗很大，但是数据中心内部的资源利用效率却不高，因此着重分析了不同种类数据中心内部所面临的成本，其中包括服务器、基础设施、电能和网络布局的成本。该文提出了几个措施用以降低数据中心的成本，包括提高数据中心的灵活性、设计有效

的算法和市场机制以合理使用服务器资源，也可以对不同地区的数据中心采用不同的服务器设置，因地制宜打造合适的数据中心。Verma et al.（2008）研究了对能源消耗敏感的设施的设计、实施和评价问题，考虑了电能消耗、迁移成本及服务水平带来的收益。该文构建了一个关注电量消耗的应用管理框架（pMapper）架构及其对应算法，解决了在保证一定的服务水平下最小化电能消耗的优化问题。Goudarzi et al.（2012）考虑了在服务提供商与需求双方签订服务水平协议下的资源分配问题，其目标是在满足服务水平约束的条件下最优化系统的能量消耗。该文使用了凸优化和动态优化的相关工具来解决此问题，并通过仿真结果验证了算法的有效性。Larumbe & Sansò（2012）认为服务器地理位置、服务器和软件性能及信息的传输方式是影响云计算服务表现的重要因素。该文综合考虑了不同层级的成本，并通过凸整数优化的方法解决了该问题。结果表明，数据中心的数量和网络建设者的预算对整个网络表现有至关重要的影响。Larumbe & Sansò（2013）设计了目标为提供高质量的服务、低成本且低排放的数据中心服务系统。该文将问题转化为混合整数规划问题并提出了一个效率很高的禁忌搜索算法，该算法可以解决有 500 个需求点、1000 个候选数据中心的规模的问题。Barroso et al.（2013）认为数据中心不能简单地被认为是一堆服务器的集合，而是应该当作一个具有仓库规模的计算机（warehouse-scale computer，WSC）来对其内部进行设计。该文介绍了此类数据中心的架构及影响其设计、运营、成本结构及软件特点的主要因素。Iyoob et al.（2013）将云计算的供应量与传统拉式供应链进行了对比，发现与传统不同的是，云计算的供应链的信息是由服务提供商流向顾客的。通过类比可以发现，云计算供应链也有类似传统供应链中的几个主体，包括服务提供商、顾客、云经纪人（cloud broker）等，该文总结了每个主体所面临的问题并为未来的研究提供了一些想法。

本部分数据中心网络设计所使用的模型是混合整数二阶锥优化问题，此类模型也已经广泛应用于各种运营管理领域。Baron et al.（2008）考虑了有随机需求和拥堵的设施选址问题，对设施的数量、位置及容量做出决策。通过分析问题的性质，作者将原问题拆分成三个子问题并设计了有效的算法。Atamtürk et al.（2012）考虑了不同的设施选址和库存管理的整合模型，并将这些模型转化成了混合整数二阶锥优化模型，文中提出

了拓展多面体割和拓展覆盖割平面，能够有效缩减优化问题的搜索空间，加快求解速度。Mak et al.（2013）考虑了可交换电池的电动汽车的电池交换站选址问题，考虑到需求的不确定性，作者将该问题构造成了分布鲁棒优化问题，并转化成为混合整数二阶锥优化问题。该文用条件风险价值（CVaR）来估计随机约束（chance constraint），并使用提出的基于二分法的算法。Kong et al.（2013）研究了医生出诊的随机预约规划问题，给定患者的数量和到达顺序，为每个患者确定其预约时间。考虑到处理患者消耗时间的不确定性，作者使用鲁棒优化的方法建立了凸锥优化问题。该文基于锥优化问题的半正定放松构造了一个接近最优解的近似算法。Mak et al.（2015）考虑了类似上述的预约计划排班问题，通过最小化最差情况下的等待时间和逾期时间，构造出了可解的锥优化问题。当已知处理时间的前两阶矩的时候，问题可以进一步转化成为二阶锥优化问题。作者通过分析问题的性质，证明了最优情况下应该按照处理时间方差的升序排列患者。Qi et al.（2015）综合考虑上网电价补贴政策（feed-in tariff）下风电的传输和存储问题，同时优化能量存储设施的容量和选址、网络结构和传输网络的承载能力。该文首先考虑无约束限制的模型，并将其构造为可解决的混合整数二阶锥优化问题，然后通过计算出显式的界来得出能量存储设备的容量。Kong et al.（2020）考虑了患者根据时间选择不履行预约的行为，使用了分布鲁棒优化模型，并转化为 copositive 规划问题，可以使用半正定规划来近似。同时，作者发现当患者的不履约行为是关于时间内生的，那么问题可以转化为两段线性的 copositve 规划问题。作者通过使用对偶价格来搜索排班计划并依次构造了近似最优的算法。He et al.（2017）研究了城市中的电动汽车共享出行问题，权衡顾客的覆盖度及运营成本。模型考虑了顾客选择行为和电动汽车的充电行为，建立了分布鲁棒优化问题，并使用混合整数二阶锥优化模型做近似处理。Sen et al.（2019）考虑了顾客选择服从混合多元逻辑（multinomial logit，MNL）模型假设下的选品优化问题。此类问题如果使用混合整数线性规划求解比较小的算例都会非常困难，但是该文采用了新颖的二阶锥优化的建模方式，并加入了 McCormick 不等式，大幅提升了问题的求解速度。

## 2.3 鲁棒动态规划相关文献

此部分主要梳理服务水平协议下的云计算服务资源供给部分中应用鲁棒动态规划方法的相关文献。在做决策的过程中，优化问题的很多参数都存在着不确定性，这些不确定的参数可能给决策的制定带来巨大影响，因此需要谨慎对待这些不确定性。其中一种解决方法就是使用鲁棒优化的方法，也就是考虑在不确定性对目标函数带来最差情况（worst case）下的最优决策。值得注意的是另一种常见的方法——随机优化（stochastic optimization）。在随机优化中，一般使用条件约束（chance constraints）来刻画不确定性，一般已知不确定参数的分布；但是对于鲁棒优化，一般是由模糊条件约束（ambiguous chance constraints）替代，并且考虑最差情况下的最优决策，此时一般只知道不确定参数的分布的部分信息。虽然鲁棒优化相对随机优化看似更加保守，但是鲁棒优化需要的信息更少，而且往往计算复杂度更低。鲁棒优化往往可以通过调整不确定集（ambiguity set）的大小来调整决策者的保守程度，而随机优化对不确定性质的度量往往很难调整。

Ben-Tal et al.（2009）对鲁棒动态规划进行了全面的介绍。动态规划面临着维度灾难的问题，而对于具有不确定性的动态规划，除了维度灾难外，还存在着不确定性灾难问题（the curse of uncertainty）。概率转移矩阵的扰动对模型的结果影响可能很大，但是现实生活中很多时候对转移概率的估计很难做到精确，尤其当转移矩阵是随时间变化的时候，决策者所面临的不确定性很大，因此解决该类问题也极具挑战性。因此鲁棒动态规划考虑有一个外界的参与者（nature）会选择在不确定集合中选择对决策者最差的转移概率、此时的决策者的最优决策。常见的不确定集合有以下几种类型：情景模型（scenario model）、间隔模型（interval model）、似然模型（likelihood model）、熵模型（entropy model）、椭球模型（ellipsoidal model）。在选择了合适的不确定集合后，就可以定义鲁棒动态规划问题。如果假设转移矩阵是关于时间独立的，那么对于有限期、有限状态及有限决策空间的问题，可以只用逆向递归（backward recursion）的方法进行求解。但是该方法只适用于状态空间、决策空间以

及决策期数数量级不大的情况。如果我们进一步考虑一些决策可以在某些不确定性的参数实现了之后再进行，换句话说，不同期的决策还依赖于前期不确定参数的实现值，那么问题会更加复杂，此类问题被称作鲁棒可调节动态规划（Robust adjustable multistage optimization）。该问题一般涉及无穷维优化问题（semi-infinite optimization），因此求解复杂度很高，一般会使用近似算法处理此类问题。

近些年，对鲁棒动态规划的研究逐渐受到了海内外学术界的重视，通过总结整理，主要可以将这些研究分为以下几个流派：

一些学者将鲁棒动态规划与普通的马尔科夫决策过程建立联系。Nilim & Ghaoui（2005）和 Iyengar（2005）是研究该领域的先驱。他们考虑一个有限状态、有限决策的马尔科夫决策过程的鲁棒控制问题，并使用非凸集描述了转移矩阵的不确定性。他们证明，当不确定集具有一定的“矩形”（rectangularity）性质时，也就是对于每个状态和行动下的概率转移矩阵和奖励都是互相独立的，那么马尔科夫决策过程中所使用的值迭代和策略迭代的方法都可以扩展到对应的鲁棒动态规划问题。Delage & Ye（2010）考虑了由均值和方差构成的不确定集合，发现此类分布式鲁棒动态凸规划问题可以转化为半正定规划问题，对于很多种目标函数都可以使用多项式时间的算法对其求解。Xu & Mannor（2012）将不确定集拓展到了分布不确定集合中，并证明了此类问题可以转换为标准的鲁棒马尔科夫决策过程，并可以使用多项式时间算法对其进行求解。Yu & Xu（2016）在 Xu & Mannor（2012）的基础上，考虑了更加一般化的分布式不确定集合，并提出了有效的算法，可以在一般的条件下很快求出最优策略。Wiesemann et al.（2013）拓展了“矩形”的概念，从对于每个状态和行动下的概率转移矩阵和奖励都是互相独立的 $(s, a)$-rectangular，拓展到了只对每个状态下的不确定集合都是独立的 $s$-rectangular。其文构造了一个策略迭代方法，在迭代过程中只需要解决中等规模的锥优化问题。Mannor et al.（2016）在 $s$-rectangular 的基础上进一步考虑了 $k$-rectangular 的不确定集合的构建。粗略来说，是根据每个状态的参数的实现值，将不确定集合投射到至多 $k$ 个不同的集合中。其思路是假设有一些状态之间可以使用相同的不确定集合，从而放松了相互独立的假设。

一些学者关注了鲁棒动态优化的最优策略的时间不一致性（time in-

consistency），也就是说在第一阶段直接计算出所有的策略并不能达到最优，相反，简单的线性调整策略在很多类型的鲁棒动态规划中的效果更好。Ben-Tal et al.（2004）考虑了一个有不确定参数的线性系统，其中一些决策可以在某些不确定参数实现后再进行决定。该文提出了线性的可调整鲁棒问题，并证明了此类问题可构造为线性规划问题或者半正定规划进行求解。Bertsimas et al.（2010）证明了关于不确定参数的线性函数控制策略在一维有约束的多阶段鲁棒优化问题中的最优性，并提出了找到此最优策略的有效算法。此方法在经典的库存控制中有着很好的表现。Bertsimas & Goyal（2012）考虑了一个两阶段的鲁棒优化问题，并给出了线性可适应控制策略最优所需要的条件，结果发现在最差情况下线性可适应策略相比完全可适应控制策略表现有可能要差上一倍。Iancu et al.（2013）放松了 Bertsimas & Goyal（2012）的假设，发现只要不确定集合满足单位超矩形中的整数次格栅（integer sublattices），而且动态规划值函数关于不确定参数的凸超模函数，那么关于不确定参数的线性决策则是最有效决策。Bertsimas et al.（2019）将线性决策规则应用到了可以使用二阶锥表示出来的分布不确定集合中，发现只要将不确定集合做一个提升（lift），并使线性决策包含关于此提升的参数，所得到的决策可以满足鲁棒动态规划问题的可行性。Hanasusanto & Kuhn（2018）考虑了 Wasserstein 距离下的不确定集合，使用线性决策法则的决策可以转化成锥优化问题，并拥有很好的近似性。

一些学者将近似动态规划和鲁棒动态规划相结合，使用静态策略来分析近似有限周期的鲁棒动态规划问题。Petrik（2012）考虑了分布不确定集合下的鲁棒近似动态规划问题，证明了鲁棒近似策略迭代算法的收敛性，此算法相比普通近似迭代算法的误差界限更小。Petrik & Subramanian（2014）在 Petrik（2012）基础上考虑可以将一些状态对应的策略合并，在保证算法的计算复杂度的情况下大幅提高解的质量。而 Lim & Autef（2019）则更进一步，考虑了如果使用核函数将状态与策略在另一空间进行投影，其计算复杂度并没有提升很多，但是解的质量得到了进一步的提升。在最近的一项仍在进行的工作中，Derman & Mannor（2020）将上述鲁棒近似算法应用在了 Wasserstein 距离下的不确定集合中，也有着不俗的算法表现。

最后一个流派的学者使用数据驱动的方法来研究鲁棒动态优化或者鲁棒增强学习问题。增强学习是机器学习中非常重要的一个分支，此流派也受到了运营管理和计算机科学等交叉学科的青睐。本流派考虑如何从数据中估计出合理的不确定集合的参数，并提出相应的算法来解决此类问题，一般来说更加关注实际的求解效果而不是其理论表现。鲁棒动态优化在库存管理（张松涛等，2015；张曙红、魏永长，2015；李春发等，2014）、供应链管理（徐家旺、黄小原，2006）、动态定价（李春发等，2014；冉伦等，2009）、网络设计（李政玲，2016；Ning & You，2019）等方面又有着重要作用。Hanasusanto & Kuhn（2013）价值函数在状态和决策上是凸二次方函数，而转移矩阵关于状态和决策是线性的。决策的可行集是二阶锥形状的，并使用 $\chi^2$ 距离构建不确定集。张松涛等（2015）研究了库存切换下的不确定动态供应链网络系统的鲁棒运作问题。针对该动态供应链网络，作者构建了参数及需求不确定的模型，并提出了模糊鲁棒控制策略。通过分析发现，该策略可以有效保证库存切换的平稳性，同时通过数值实验验证了提出的模糊鲁棒控制策略的有效性。张曙红、魏永长（2015）考虑了逆向物流下的供应链中的鲁棒动态库存模型，使用鲁棒优化的方法分析了该闭环供应链的动态性能。通过仿真可以发现，将制造与再制造流程综合起来的鲁棒控制策略能够减小库存波动，从而降低库存成本。Yang（2017）考虑了有限周期的鲁棒动态控制问题，并使用 Wasserstein 度量对不确定集进行建模，证明 Markov 策略的存在性和最优性，并开发基于凸优化的算法来计算和分析最优策略。陈美蓉等（2017）提出了鲁棒动态多目标优化问题的优化方法，定义了性能鲁棒性和时间鲁棒性，对多目标问题进行分解。该文使用了移动平均的预测模型来预测时间序列，并对两个鲁棒性的评价测度进行了仿真实验，证明了方法的有效性。Ning & You（2019）考虑了农业的废弃物能源转化网络的鲁棒优化问题，作者构建了一个两阶段的鲁棒动态规划问题，并且使用 Wasserstein 度量构建不确定集合，结果发现该方法下的平均成本要比传统的随机优化得到的结果低 5%以上。

为了更好地总结归纳相关文献，本书对鲁棒动态规划相关的代表性文献的整理见附录 A 表 A-1。

## 2.4 结 论

本书聚焦于新能源汽车供应链与数据中心供应链这类新兴技术下的供应链所面临的资源分配管理问题。其中包括了数据中心的选址和服务资源分配、电动汽车补贴资源分配和云计算备用服务器的资源分配问题。综合使用了管理科学中的整数规划、博弈论、鲁棒优化和动态规划等工具对问题进行了建模和求解。

新能源汽车补贴资源分配部分总结了影响消费者购买电动汽车意愿的实证类型的文章，为消费者的效用模型奠定了基础，同时也总结了使用实证和仿真等方法对政府补贴政策的效果估计。最后汇总了使用博弈等分析模型研究政府的最优补贴政策的文章，为此部分研究的建模打下基础。

数据中心供应链网络布局与资源分配部分总结了数据中心选址与传统供应链选址的相似与不同，并将数据中心的网络设计问题归结为 SLCIS 问题。在解决此类问题时，通常需要解决规模较大的混合整数锥优化问题，因此文献综述部分也总结了使用类似模型的应用型文章及关于解决此类问题的方法类型的文章。

最后重点梳理了近些年与云计算资源动态调整密切相关的鲁棒动态规划相关文献。主要的文献可以分为四类，分别是不确定集合的矩形性质对动态规划求解的影响、线性决策规则的最优性、近似动态规划与鲁棒动态规划的结合及数据驱动下的鲁棒增强学习。鲁棒优化作为管理科学学科近些年备受关注的工具在服务资源分配的场景中有着很大的应用潜力。

综上所述，供应链的运营管理和资源分配问题涉及了管理科学学科中不同研究工具，每个子领域在近些年都受到了学术界和业界的高度重视，对这些问题的研究不仅有助于拓展学术前沿，也具有影响深远的实际意义。

# 第 3 章　新能源汽车供应链的补贴资源分配

本部分以新能源汽车供应链为例，主要研究电动汽车市场的补贴资源分配问题。电动汽车能够有效减少温室气体的排放，但是其作为新兴技术仍未被大众所接受。各地政府采取了相应的激励手段以促进电动汽车行业的发展，一方面可以将补贴直接提供给消费者，另一方面也可以补贴充电基础设施建设。相比过去的以计量模型为主的研究，本部分创新地考虑了充电桩与电动汽车之间的正向网络效应，并建立分析模型计算出政府最优的补贴资源分配策略，分析了科技进步、市场特点等因素的变化将如何影响最优补贴分配政策。最后还搜集使用了深圳市的相关数据，验证了同时补贴消费者和充电桩建设的混合补贴政策的优势。

## 3.1　引　　言

空气污染一直是全球范围内的重要环境问题。世界卫生组织（WHO）强调空气污染是人类健康的最大环境风险，据估计空气污染每年造成 700 万人过早死亡。不幸的是，91%的世界人口居住在空气质量低于 WHO 标准的地区。[①] 根据美国环境保护局（EPA）的估计，在美国，机动车造成了总一氧化碳污染的 75%。据估计，在美国，道路车辆造成的空气污染占 1/3，带来了 27%的温室气体排放。[②]

电动汽车被认为是减少空气污染和温室气体排放的解决方案之一。尽管从地球上提取和处理矿物来制造电动汽车电池可导致碳排放，但最近的研究表明，在整个生命周期中，使用电动力和内燃机的汽车的空气污染排放差异

① 世界卫生组织，网址为 https://www.who.int/airpollution/en/。

② https://auto.howstuffworks.com/air-pollution-from-cars.htm

很大。由于没有燃烧和排气管排放，电动汽车比汽油和柴油驱动的汽车更具优势。目前，越来越多的电动汽车被采用，并且电池回收技术日趋成熟，对新电池的需求将越来越小，电动汽车对环境更加有益。[①]电动汽车还可以帮助各国减少对国内或国外石油资源的依赖，避免油价波动带来的经济震荡等后果。

由于电动汽车是高科技创新产品，因此消费者可能会对电动汽车持保守想法，包括充电基础设施不足、续航距离短、购买价格高或潜在的电池问题（Zhang & Bai，2017）。如果没有政府的适当干预，电动汽车市场可能面临严重的自我萎缩问题（critical mass problem），也就是说由于市场接受率无法自我维持并实现进一步的增长，从长远来看，电动汽车市场最终将萎缩甚至消失（Zhou & Li，2018）。消费者方面的这些担忧以及电动汽车的环境利益促使世界各国政府提出支持计划以促进电动汽车的采用。美国、西班牙、德国、挪威和中国等国家已经启动了支持计划，以鼓励消费者购买电动汽车。在美国，联邦政府和一些州政府为购买电动汽车提供补贴（Holland et al.，2016）。西班牙政府为电动汽车的价格提供25%的折扣（Luo et al.，2014a）。

除了直接补贴电动汽车消费者之外，一些国家也提供了其他支持性项目来推广电动汽车。挪威政府在补贴电动汽车消费者的同时，也会为建造充电桩提供经济支持（Springel，2016）。支持建造充电桩有利于减轻顾客的里程焦虑。所谓里程焦虑，指的是一些消费者对电动汽车的行驶距离不能满足其要求，而又难以找到充电桩充电，以致抛锚在路的担心（Lim et al.，2015）。德国政府在 2016 年启动了一个价值 10 亿欧元的电动汽车补贴政策，为电动汽车的购买者提供税豁免并为充电桩的建设提供 30%的补贴（Zhang & Dou，2020）。中国在 2009 年就提出了电动汽车的补贴计划，并在 2013 年对计划进行了升级。该计划也是世界上最具魄力的计划之一（Hao et al.，2014）。中国政府针对不同方面提供了支持，具体可分为以下三个方面：（1）对购买电动汽车的补贴；（2）针对基础设施的建设和补贴；（3）对电动汽车研发的投资（Zhang & Bai，2017；Zhang et al.，2017）。

① https://www.forbes.com/sites/jamesellsmoor/2019/05/20/are-electric-vehicles-really-better-for-the-environment/#30f0728b76d2

尽管各国政府都付出了不少努力，但电动汽车的接受率仍与期望有一定差距。例如，在 2010 年，德国政府定下了在 2020 年电动汽车保有量达到 100 万辆的目标，然而截至 2019 年 10 月，仅完成推广了 142805 辆电动汽车，与目标相比远远不足。[①] 中国的电动汽车保有量世界领先，但是电动汽车的市场占有率仍然处于较低水平。截至 2015 年 11 月，中国电动汽车的总产量刚刚超过了总汽车产量的 1%，而 2016 年和 2017 年的电动汽车销量分别为 507000 辆和 777000 辆，也仅仅占据总销量的 1.81% 和 2.69%（Zhang & Qin，2018）。深圳是中国电动汽车保有量最多的城市，在 2009 年到 2012 年启动了电动汽车补贴项目，拨款 20 亿人民币计划将电动汽车和充电桩的数量增加至 15000 辆和 12750 座。[②] 然而，在 2013 年年初，深圳仅仅推广了 2273 辆电动车和 600 个充电桩。[③] 深圳政府在其年度报告中阐明，充电桩的增长速度无法满足电动汽车车主的需求，可能会对电动汽车的推广带来负面效用。[④]

虽然电动汽车较低的市场占有率可能由多种原因造成，但缺乏有效的支持机制可能是其中最重要的原因之一。图 3-1（a）展示了深圳的电动汽车总数量和充电桩总数量，图 3-1（b）展示了对消费者和充电桩的补贴。深圳市政府于 2010 年开始补贴电动汽车消费者，但直到 2014 年才向充电桩建设提供财政支持。尽管 2014 年及以后几年电动汽车消费者的补贴有所减少，但可以发现电动汽车消费者的数量在急剧增长，这表明增加对新充电桩的投资补贴确实有促进电动汽车的销量的作用。

本书旨在研究政府提供的最佳补贴资源分配策略，以实现推广电动汽车的最大净收益。政府既可以向购买电动汽车的消费者提供补贴，也可以补贴建设充电桩的成本。综合考虑电动汽车对社会的利益和补贴支付的成本后，可以构建出政府的目标函数。在电动汽车的推广过程中，充电桩和电动汽车作为互补产品形成了正反馈效应。一方面，充电桩的更高可用性减轻了消费者的行驶里程焦虑，并增加了购买电动汽车的吸引力。换句话说，充电桩为电动汽车消费者创造了积极的网络外部性，这是在设

① 数据来源：https://www.kba.de。

② 中国科学技术部，网址为 http://www.most.gov.cn/kjbgz/200909/t20090930_73529.htm。

③ 网易汽车，网址为 http://auto.163.com/13/0325/10/8QQ95PI900084IJS.html。

④ 深圳政府网，网址为 http://www.sz.gov.cn/szzt2010/zdlyzl/sj/201808/t20180829_14044796.htm。

计电动汽车支持计划时必须考虑的重要特征（Springel，2016）。另一方面，更多的电动汽车消费者增加了充电桩的利润，并使对充电桩的投资更具吸引力。模型考虑了充电桩和电动汽车之间的这种相互作用。消费者购买电动汽车的效用随着充电桩数量的增加而增加，每个充电桩的利润随着电动汽车数量的增加而增加。本部分计算了政府的最佳补贴金额，并研究了模型参数对补贴金额及市场上电动汽车和充电桩数量的影响。综上，本部分的主要研究结果可以为政策制定者提供如下建议：

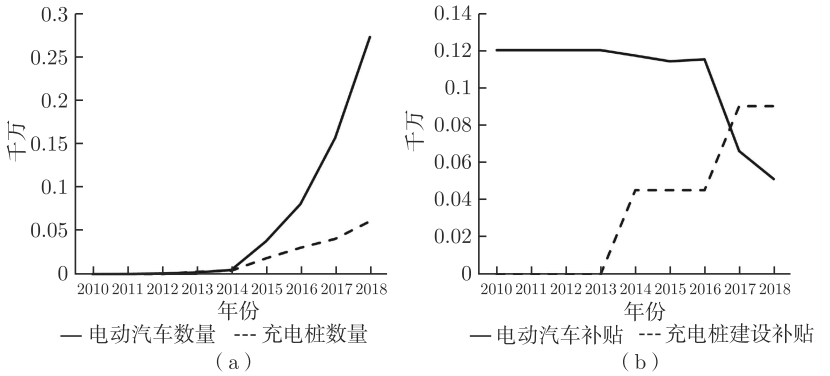



**图 3-1　深圳市电动汽车、充电桩数量以及补贴政策**

（1）如果电动汽车的社会效益上升，那么当充电桩对电动汽车消费者的效用产生更强的网络效应时，政府应更多地向充电桩提供补贴，而对电动汽车消费者的补贴则可减少。但是，如果网络效应不强，政府应增加对消费者的补贴或减少对充电桩的补贴。如果政府仅补贴电动汽车消费者时，情况则有所不同。此时，当电动汽车变得对社会更有利时，政府总是会增加消费者补贴。

（2）直观上来讲，当建造充电桩的成本增加时，政府应始终增加对充电桩的补贴。但是，经过推导证明可以发现这种策略并不总是最佳的。实际上当消费者足够重视电动汽车的环境效益，且电动汽车的成本不太高时，政府应增加对投资者的补贴额；否则政府应实际减少对充电桩的补贴。

（3）直观上来说，政府在消费者的绿色环保意识加强时，应该减少对消费者的补贴，增加对投资者的补贴。然而，情况并非总是如此，消费者

的环保意识对补贴的影响取决于充电桩的成本。更具体地讲，如果建造充电桩的成本高昂，则随着消费者对电动汽车绿色程度的认可度的增加，政府应更多补贴充电桩或减少补贴消费者。另一方面，如果建设充电桩的成本很低，随着消费者对电动汽车绿色程度的评估的增加，政府应减少对充电桩的补贴。关于消费者补贴，政府最初增加了补贴，并利用消费者对电动汽车环境效益的欣赏，促进电动汽车的普及。但是，当消费者的环保意识足够高时，政府将减少消费者补贴，不再需要在补贴上花费更多。

（4）本部分将模型与政府仅补贴电动汽车消费者的模型进行了比较。有趣的是，当充电站的成本很高时，政府反而无须补贴充电桩的投资者。政府只要支持电动汽车消费者就可以实现相同数量的电动汽车和充电桩。然后，运用来自深圳的数据对模型进行了校准，发现同时补贴电动汽车消费者和充电桩投资者比仅补贴消费者具有明显的优势。联合补贴政策下的电动汽车和充电桩数量可以达到单补贴政策下的 3 倍和 27 倍。此外，政府在联合补贴政策下的净收益平均为单一补贴政策下净收益的 2 倍，最高为 5.74 倍。

本章的结构如下：3.2 节回顾了文献；3.3 节介绍了使用的符号以及模型框架；3.4 节刻画了最优补贴政策，并比较了单边补贴和混合补贴政策。3.5 节探究了不同科技水平和市场特点下的政府补贴政策变化。3.6 节使用来自深圳的电动汽车数据对模型进行校准；3.7 节总结本章。

## 3.2 文 献 综 述

本部分的工作属于运营管理中的电动汽车问题相关研究。一些论文使用实证模型来预测电动汽车补贴政策的有效性。Hao et al.（2014）、Zhang & Bai（2017）和 Zhang et al.（2017）回顾并总结了中国的电动汽车政策，并估计了补贴结束后的电动汽车市场规模。Mueller & Haan（2009）使用了基于 agent 的仿真模型来预测补贴政策的效果。类似的，Gnann et al.（2015）也采用了类似的仿真模型来拟合德国的电动汽车市场发展状况。

一些文献采用分析模型来研究电动汽车问题。Lim et al.（2015）研究了里程焦虑和转售焦虑对电动汽车发展的影响。他们发现，当转售焦虑

程度分别为低或者高时，拥有或者租赁电池与增强充电服务的组合通常会在电动汽车采用的目标（排放节省、盈利能力和消费者剩余）之间达到最佳平衡。Huang et al.（2013）分析了政府补贴电动汽车消费者时燃料汽车供应链与电动燃料供应链之间的竞争。他们发现，更多的服务和充电桩可以增加补贴对电动汽车市场的积极影响；但是，模型中的充电桩的数量是其模型中的一个参数，并没有分析最优的充电桩的补贴。Luo et al.（2014a）考虑一种分权管理下的电动汽车供应链，其中政府为电动汽车消费者提供价格折扣，并且有一定的补贴上限。该研究表明，当单位生产成本较高时，补贴上限对影响制造商的最佳批发价格更为有效，而在生产成本较低时，贴现率则更为有效。Cohen et al.（2015）研究了政府在需求不确定的情况下，在两个时期内向分散的供应链提供绿色产品补贴的模型。政府可以承诺补贴或保留在各个时期内调整补贴的选择。该研究表明，灵活的补贴政策平均而言要昂贵得多，除非各个时期之间存在显著的负需求相关性。另一方面，政府在弹性环境中的额外支出减少了电动汽车推广水平的不确定性。Shao et al.（2017）假设政府为电动汽车消费者提供固定补贴或价格折扣，分析了垄断和双头垄断下的电动汽车和汽油汽车市场。该研究表明，这两种激励方案对环境和社会福利的影响相同，但由于支出较低，政府更喜欢采用补贴机制。

文献中仅有少数一些研究考虑了充电桩与电动汽车的相互作用。Yu et al.（2016）构建了一个有电动汽车和充电桩的双边市场序贯博弈模型，并计算出其均衡。这些表明，社会最优的解决方案需要比市场竞争下建设更多的充电桩。尽管作者在模型参数的敏感性分析中谈到了补贴的效果，但他们并未分析政府的最优补贴政策。Springel （2016）使用结构模型来分析电动汽车和充电桩的双边市场。该文利用来自挪威的数据，提供了描述性证据，证明电动汽车的购买与消费者补贴和充电桩补贴均呈正相关。Zhou & Li（2018）考虑到充电站部署和电动汽车采用之间的相互依赖性。他们使用美国大都市统计区（MSA）的面板数据发现，超过一半的 MSA 面临自我萎缩问题（critical mass problem），补贴政策可能会更有效地促进电动汽车的普及。无论是 Springel（2016）还是 Zhou & Li（2018）都没有考虑政府的最优补贴政策。

Zhang & Dou（2020）是最接近本部分工作的论文。他们考虑了一个

充电桩投资者（服务提供商）必须根据以下情况折中选择充电桩位置的问题：在城市地区建造充电桩较为昂贵，但由于交通流量较大，可能产生更多的收入；而充电桩在郊区建造，房屋成本较低，但使用频率也较低。该文表明，在平衡状态下，充电桩的需求和供应与电动汽车的普及之间可能存在空间失配。他们研究了政府应如何使用三项补贴政策来缓解这种失衡：（1）按照充电桩数量补贴投资者；（2）按照充电桩使用率来补贴投资者；（3）补贴电动汽车消费者。文章主要发现，通过使用电动汽车来补贴电动汽车消费者或补贴服务提供商可能是政府的最佳补贴策略，而且在许多情况下，政府不应鼓励服务提供商在城市地区投入更多的电动汽车。尽管 Zhang & Dou（2020）分析了政府补贴，但他们专注于理解单一补贴如何影响单一服务提供商在两个不同位置（城市与郊区）建造充电桩的决定。而与其不同的是，本研究主要考虑电动汽车与充电桩之间的互补性，并将重点关注政府应如何同时向电动汽车消费者和充电桩投资者提供补贴，分配补贴资源。

本书也与互补产品以及网络效用相关的文献有所联系，比如 Katz & Shapiro（1985）、Farrell & Saloner（1985）、Caillaud & Jullien（2003）、Rochet & Tirole（2006）、Armstrong（2006）。这些文章主要聚焦于双边市场中的定价与协调策略。在运营管理和营销领域，Bhaskaran & Gilbert（2005）研究了耐用品制造商的租用与售卖策略在有其他互补产品时的影响。Yalcin et al.（2013）研究了两个生产互补产品的企业的定价和质量策略。He et al.（2016）研究了产品和顾客的特点会如何影响互补企业的团购决策。He & Yin （2015）研究了供应链中的竞争会如何影响互补供应商的销售策略。这些文章没有站在政府的角度，为互补产品提供最优补贴政策的设计。

总之，现有文献大多使用实证模型来估计电动汽车的推广水平。而使用分析模型的电动汽车相关论文，要么将补贴作为模型参数，不分析政府的问题，要么仅考虑一种补贴而忽略了网络效应。相比之下，本部分提出了一种优化模型，该模型同时包含了针对电动汽车消费者和充电桩投资者的补贴。通过计算政府的最佳补贴政策，并进行敏感性分析，以了解不同参数对最优政策的影响。

## 3.3 模　　型

政府希望通过其政策来促进电动汽车的推广。为了实现这样的目标，政府可以使用现实生活中常见的两种不同的补贴策略来激励消费者。

（1）补贴电动汽车消费者。在此策略下，电动汽车的消费者会享受到购买补贴。此策略可以使得电动汽车的价格更加亲民，提高其接受度。

（2）补贴公共充电设施。此补贴策略主要针对潜在的投资者，减轻公共充电桩的投资成本。通过这样的补贴，可以吸引更多的充电设施的建设，从而减轻电动汽车司机的里程焦虑，加强了电动汽车相对于传统汽车的吸引力。

将此问题建模成一个斯达克伯格博弈，首先政府公布其补贴策略，接下来，顾客和投资者同时进行其购买和投资决策，同时对对方的决策产生理性预期。在设计补贴政策的时候，政府的目标是最大化电动汽车推广对社会的净收益，同时考虑了更高推广带来的社会收益以及补贴带来的支出。接下来描述消费者和投资者的决策过程，并刻画其最优反应。最后建立政府的优化问题，并给出最优补贴政策。

### 3.3.1 消费者效用模型

每个消费者通过最大化个人效用来决定其购买策略。电动汽车的消费者可能是当前的电动汽车司机，也可能是未来的潜在消费者。分别用 $\lambda_0$ 和 $\Lambda$ 来表示这两个群体的数量，其中 $\lambda_0$ 表示在引入补贴之前的电动汽车使用者数量，$\Lambda$ 表示该补贴政策的潜在受益者。为了刻画潜在消费者的决策行为，接下来构造了其效用函数。

目前关于电动汽车实证模型的文章，例如 Lin & Wu（2018）、Han et al.（2017）、Degirmenci & Breitner（2017），指出顾客是否选择购买电动汽车要考虑多种因素，包括经济因素、使用便利性、环保意识等。从经济角度来看，消费者需要承担电动汽车的零售价格、政府对消费者的潜在补贴及驾驶电动汽车的成本（即充电费）。电动汽车对消费者的吸引力还取决于他们在使用电动汽车时遇到的不便和障碍，这主要是由于缺乏充电桩的可用性。鉴于电动汽车的行驶里程有限（与传统汽油车相比），充

电基础设施可以减轻消费者的里程焦虑，并使电动汽车成为更可行的选择（Avci et al.，2014；Lim et al.，2015）。此外，消费者是否选择电动汽车还取决于他们对保护环境的态度，这种态度因消费者群体而异。

综合考虑这些重要因素，假设顾客关于环保态度是异质的，对于类型为 $\theta$ 的消费者，其效用见式 (3-1)：

$$u_\theta(m|s) = \theta v - \mu(m) - (p - s) - \phi \tag{3-1}$$

其中 $v$ 表示顾客获得电动汽车带来的单位名义效用，$v$ 与顾客类型 $\theta$ 表示不同顾客对环境保护的态度下的实际效用。假设 $\theta$ 在 $[0, \bar{\theta}]$ 上均匀分布，其中 $\bar{\theta}$ 表示最环保的顾客。式 (3-1) 中的 $m$ 表示电动汽车消费者可用的公共充电桩数量，而 $\mu(m)$ 表示由充电桩不足造成的里程焦虑。假设 $\mu(m)$ 是一个关于 $m$ 单调递减的凸函数，这样多建造一个充电桩可以减小一定里程焦虑，但是其效果边际递减。参数 $\phi$ 表示平均的总充电费用，可以认为 $\phi$ 是电动汽车使用期间的总行驶距离乘以每千米的充电费用。最后 $p$ 和 $s$ 分别表示了电动汽车的价格及购买补贴，因此 $p - s$ 为消费者需要实际付出的购买费用。

给定政府的购买补贴 $s$、充电桩的数量 $m$，类型为 $\theta$ 的消费者比较其购买电动汽车的效用 $u_\theta(m|s)$ 与其他选择的利润 $u_0$（比如购买传统汽车，或者什么也不购买），并选择可以为其带来更高效用的决策。因此，存在一个阈值 $\theta_1(m|s) \in [0, \bar{\theta}]$，只要顾客的类型满足 $\theta \geqslant \theta_1(m|s)$，这个顾客就会选择购买电动汽车。因此，给定补贴时，电动汽车的数量满足式 (3-2)：

$$\begin{aligned}\lambda(m|s) &= \lambda_0 + \Lambda \Pr\left(u_\theta(m|s) \geqslant u_0\right) = \lambda_0 + \Lambda \Pr\left(\theta \geqslant \theta_1(m|s)\right) \\ &= \lambda_0 + \max\left\{0, \left[\frac{v\bar{\theta} - u_0 - (p - s) - \phi - \mu(m)}{v\bar{\theta}}\right]\Lambda\right\}\end{aligned} \tag{3-2}$$

### 3.3.2 充电桩投资者决策

潜在的投资者会根据是否有利可图来决定是否进入市场投资建造充电桩。假设每个充电桩都是同质的，并且平均分配从消费者那里赚取的利润，也就是说充电桩市场面临着充分竞争。因此如果增加一个充电桩能够有利润，那么就会有新的投资进入。展开来说，如果潜在的投资者投资充电桩的收益为 $\pi$，其外部选择的潜在收益为 $\pi_0$，那么这个投资者会进入市场，当且仅当 $\pi \geqslant \pi_0$，其中 [见式 (3-3)]：

$$\pi(\lambda,\ m|\kappa)=\frac{\lambda\phi}{m+1}-(f-\kappa) \tag{3-3}$$

式 (3-3) 中的第一项计算了给定当前电动汽车数量 $\lambda$ 和充电桩数量 $m$ 的时候，新进入的充电桩的总收益。具体来说，充电桩的总收益来自电动汽车司机的总充电成本 $\lambda\phi$，这些收益平均分到了所有运营的充电桩中（Springel，2016）。这里假设了充电费用是外生的，而非充电桩运营商内生决定。这是因为关于充电费用的制定，有相关法律的规定加以限制。比如北京地区的充电费用一般由电费和充电服务费构成，其中充电服务费每千瓦·时收费上限为当天北京 92 号汽油每升最高零售价的 15%。[①] 因此，式 (3-3) 表示了第 $m+1$ 个充电桩进入市场后的期望收益。$f$ 表示建造一个充电桩的固定成本，$\kappa$ 表示政府提供给充电桩投资者的补贴。因此，给定电动汽车市场规模 $\lambda$、政府的投资补贴 $\kappa$，此时均衡条件下的充电桩投资者市场规模 $m(\lambda|\kappa)$ 满足式 (3-4)：

$$m(\lambda|\kappa)=\max\left\{m_0,\ \frac{\lambda\phi}{\pi_0+f-\kappa}\right\} \tag{3-4}$$

其中 $m_0$ 为补贴政策前的充电桩初始数量。

### 3.3.3　政府决策问题

政府的目标是通过权衡两种补贴策略，最大化推广电动汽车带来的社会净收益。使用 $\beta$ 表示每个电动汽车带来的社会收益，这代表了由于使用电动汽车而非传统汽车带来的环境、健康方面的提升，同时也表示了使用可持续能源带来的对国外能源依赖的减少。

预期到式(3-2) 和式(3-4)中消费者和投资者的行为，政府通过优化下面双补贴政策优化模型（TSSP）进行决策 [见式 (3-5)、式 (3-6)、式 (3-7)]：

$$\begin{aligned}\max_{s,\ \kappa}\quad & \Pi(\lambda_0,\ m_0)=\beta\lambda-s\,(\lambda-\lambda_0)-\kappa\,(m-m_0)\\ \text{s.t.}\quad & \lambda=\lambda_0+\max\left\{0,\ \left[\frac{v\bar{\theta}-u_0-(p-s)-\phi-\mu(m)}{v\bar{\theta}}\right]\Lambda\right\}\end{aligned} \tag{3-5}$$

$$m=\max\left\{m_0,\ \frac{\lambda\phi}{\pi_0+f-\kappa}\right\} \tag{3-6}$$

① 中国汽车工业协会，网址为 http://www.caam.org.cn/chn/9/cate_107/con_5167189.html。

$$0 \leqslant s \leqslant p, \quad 0 \leqslant \kappa \leqslant \pi_0 + f \tag{3-7}$$

政府目标函数的第一项表示了电动汽车推广带来的社会收益，第二项和第三项分别表示支付给消费者和投资者的总补贴支出。该优化问题的约束条件为消费者和投资者在给定补贴政策下的市场均衡数量。将在下一节分析此问题并研究均衡下的补贴策略。为了在符号上更清楚表示，定义 $\hat{\theta} = v\bar{\theta}$，$c = \pi_0 + f$，其中 $\hat{\theta}$ 可以表示为顾客对电动汽车的最大估值，而 $c$ 表示了投资者进入市场的实际壁垒（详见表 3-1）。

**表 3-1 本章主要符号列表**

| | |
|---|---|
| $v$ | 消费者对新车的估值 |
| $p$ | 电动汽车售价 |
| $\phi$ | 平均充电费用 |
| $\theta$ | 消费者环保意识强度，服从 $[0, \bar{\theta}]$ 的均匀分布 |
| $u_0$ | 消费者的外部效用 |
| $\Lambda$ | 潜在电动汽车消费者数量 |
| $\lambda_0$ | 当前电动汽车持有者数量 |
| $f$ | 充电桩的投资建造成本 |
| $\pi_0$ | 投资者的外部效用 |
| $s$ | 给每个电动汽车消费者的补贴 |
| $\kappa$ | 给每个充电桩投资者的补贴 |
| $\beta$ | 电动汽车的使用带来的社会收益 |
| $\lambda$ | 补贴后使用电动汽车的消费者数量 |
| $m$ | 补贴后公共充电桩的数量 |
| $\widehat{\theta} = v\bar{\theta}$，$c = \pi_0 + f$ | |

为了模型的可解性，本部分的分析主要聚焦在公共充电桩初始数量 $m_0$ 比较小的时候。该假设可以帮助我们得到显示解，而且对于大部分电动汽车市场尚不成熟的地区而言，也非常贴近现实。不仅如此，对于一些发达地区和一些已经提供一段时间补贴的地区而言，也有类似的现象。根据《纽约时报》和《福布斯》的报道，美国的电动汽车充电桩还远远不够，全美缺乏公共和家用电动汽车充电桩是购买电动汽车的主要障

碍。[①] 当中国在 2010 年开始其电动汽车补贴政策时，全国只有 76 个公共充电站；[②]挪威在 2009 年刚开始进行补贴项目的时候，也只有不超过 200 个充电点（Kvisle，2012）。其他国家和地区，比如澳大利亚、俄罗斯、印度等的情况也类似。[③] 这样的假设意味着电动汽车的拥有者在初期只能依靠家用和工作场所的充电桩进行充电，很少有公共的充电桩供其使用。为了研究的完整起见，本部分也在附录中对更为一般的 $m > 0$ 进行了讨论，刻画了此情况下的最优补贴政策。虽然其最优补贴政策的结构非常复杂，难以直接进行分析，但是通过数值实验证实了本部分的发现在初始公共充电站数量不为零的时候依然成立。

## 3.4　模 型 分 析

为了得到政府的最优决策，首先分析式(3-2) 和式(3-4)在给定补贴 $(s, \kappa)$ 的均衡市场状况，然后将此均衡代入政府的最优化问题中，并求解出 TSSP 问题的最优解。首先做出以下假设：

**假设 3-1：** 分析中，假设

（1）$\widehat{\theta} > u_0 + p + \phi + \mu(0)$；

（2）$c > \Lambda r^2/4\widehat{\theta}$。

该假设的第一部分确保即使在没有任何政府支持的情况下，总有消费者愿意购买电动汽车。第二个不等式确保用于电动汽车购买者和投资者的最佳补贴值分别不超过电动汽车价格和投资成本（即约束 $0 \leqslant s \leqslant p$，$0 \leqslant \kappa \leqslant c$ 没有达到上限）。这两个假设都是为了避免缺乏现实意义的边界解。

---

① 《纽约时报》，网址为 https://www.nytimes.com/2020/04/16/business/electric-cars-cities-chargers.html，《福布斯》，网址为 https://www.forbes.com/sites/brookecrothers/2019/10/13/in-the-us-electric-vehicle-charging-prospects-are-bleak-out-there-for-the-rest-of-us-who-dont-drive-a-tesla-model-3/#172a5b8533d1。

② https://www.prnewswire.com/news-releases/china-ev-charging-station-and-charging-pile-market-2018-2025-15-global-and-chinese-charging-operators–operation-and-development-strategies-of-8-chinese-ssuppliers-300701521.html

③ 澳大利亚，https://www.caradvice.com.au/830862/electric-car-fast-charging-network-priority-australia/，俄罗斯，https://blogs.platts.com/2018/03/26/russia-electric-vehicles-ev/，印度，Nair et al.（2017）。

**引理 3-1：** 对于给定的补贴策略 $(s，\kappa)$ 和起始电动汽车数量 $\lambda_0$，消费者和投资者的市场均衡是唯一的，且如下所示：

$$\widehat{\lambda} = F^{-1}\left(\lambda_0 + \left[\frac{\widehat{\theta} - u_0 - (p-s) - \phi}{\widehat{\theta}}\right]\Lambda\right)$$
$$\widehat{m} = \widehat{\lambda}\phi/(c-\kappa)\ ,$$

其中 $F(\cdot)$ 为

$$F(x) = x + \frac{\Lambda}{\widehat{\theta}}\,\mu\left(\frac{x\phi}{c-\kappa}\right)$$

是一个关于 $x$ 单调递增的函数。

相关的证明可参见附录 B。上述引理 3-1 刻画了给定政府补贴后的电动汽车市场变化。可以直观地看出，$\widehat{\lambda}$ 和 $\widehat{m}$ 关于 $s$ 和 $\kappa$ 都是单调递增的，因此更高的补贴可以通过网络效应让充电桩以及电动汽车消费者的数量都有所增长。然而，这些关系的具体性质更加复杂，政府可以利用它们来实现社会效益的最大化。

接下来，对充电桩带来的网络效应、里程焦虑的减弱进行了刻画，里程焦虑函数为 $\mu(\cdot)$：

**假设 3-2：** $\mu(m) = r_0 - r\sqrt{m}$。

采用如上形式的假设有以下几点优势。首先，可以直接求解出显示解，并且能够保证网络效应的特点。其次，其参数易于解释，可以通过其对均衡结果的影响得到相关的政策方面的启示。具体地说，$\mu(m)$ 是关于 $m$ 的单调递减的凸函数。这与传统文献中基础设施的网络影响是一致的：增加站点数量可以减少里程焦虑，但是降低速度有边际递减的性质（Jiwattanakulpaisarn et al.，2012）。此外，可以将行驶里程限制解释为电动汽车无须充电即可行驶的区域的半径。因此，随着站点数量的增加和每个特定站点的覆盖范围的相应缩小，此时的电动汽车的行驶范围会与站点数量的平方根成正比。[①] 值得注意的是，由于电动汽车行驶距离的限制，充电桩为消费者带来的里程焦虑缓解通常无法辐射到全国区域，因此本书的模型提出的补贴政策主要适用于省或者城市的范围。

① 这里假设城市中的充电桩均匀分布。假设有一个矩形的城市，居民均匀分布在城市中。如果只有一个充电桩覆盖整个区域，城市中的居民到达城市中心点进行充电。如果有 4 个充电桩均匀分布在城市中，则可以划分出 4 个相同大小的区域，其中每个区域的居民只需要去对应的充电桩充电即可。这样一来，每个居民的平均行驶距离将减半。这意味着充电桩数量每增加 4 倍，其平均行驶距离就会减小一半。换言之，居民的里程焦虑程度可以被认为是随着充电桩数量的增加以平方根的速度下降。

在假设 3-2 中，$r_0$ 表示在没有公共充电桩时（即当消费者完全依靠家庭或工作场所充电时），行驶里程的限制导致用户效用的最大减少。另一方面，$r$ 表示了公共充电桩产生的（正）网络外部性的强度。也就是说，随着更多的充电桩启动，它量化了减轻里程焦虑的速度（从而提高了用户的使用效率）。需注意到，在该定义中 $\mu(m)$ 应为非负值，因为它表示驾驶员由于无法使用充电桩而带来的里程焦虑。因此，要求 $r_0$ 与 $r$ 相比要足够大，以使 $\mu(m)$ 始终保持非负值，保证此假设的一个充分条件是 $r_0/r \geqslant \sqrt{\Lambda(\beta+\phi)/c}$。

做计算政府最优政策的第一步，考虑以下的关于 $z$ 的多项式方程：

$$4c^2\widehat{\theta}z^3 - 3\Lambda rc\phi z^2 + (2\Lambda\phi c(\phi-\beta+(p+u_0+r_0)-\widehat{\theta})- 4c\phi\widehat{\theta}\lambda_0)z + \Lambda r\phi^2\lambda_0 = 0$$

使用 Descartes 规则可知此方程有两个正实根。使用 $z_0$ 表示较大的实根，那么 TSSP 问题的最优解如下所示。

**命题 3-1：** 定义 [见式 (3-8)]

$$\begin{aligned}
\bar{s}_0 &= \beta+\phi-\frac{2c\widehat{\theta}(\beta+\widehat{\theta}-(p+u_0+r_0))}{4c\widehat{\theta}-\Lambda r^2},\\
\bar{\kappa}_0 &= c-\frac{2\phi c\,(4c\widehat{\theta}-\Lambda r^2)}{(\beta+\widehat{\theta}-(p+u_0+r_0))\Lambda r^2}-\frac{\phi(4c\widehat{\theta}-\Lambda r^2)^2}{\left[(\beta+\widehat{\theta}-(p+u_0+r_0))\Lambda r\right]^2}\lambda_0
\end{aligned} \tag{3-8}$$

如果 $\bar{s}_0$，$\bar{\kappa}_0>0$，那么给予消费者和投资者的最优补贴分别是 $s^*=\bar{s}_0$ 和 $\kappa^*=\bar{\kappa}_0$。否则有

$$s^*=\begin{cases}\dfrac{cz_0^2\widehat{\theta}-\Lambda r\phi z_0+\Lambda\phi((p+u_0+r_0)+\phi-\widehat{\theta})-\lambda_0\widehat{\theta}\phi}{\Lambda\phi}, & \text{if } \bar{s}_0\geqslant 0,\ \bar{\kappa}_0<0,\\ 0, & \text{if } \bar{s}_0<0,\end{cases}$$

$$\kappa^*=\begin{cases}\dfrac{c\left[(\phi^2+\beta^2)\Lambda^2r^2+4\Lambda c\phi\widehat{\theta}(\phi+(p+u_0+r_0)-\widehat{\theta})-4\phi\hat{\theta}^2c\lambda_0\right]}{(\Lambda r(\phi+\beta))^2}, & \text{if } \bar{s}_0<0,\ \bar{\kappa}_0\geqslant 0,\\ 0, & \text{if } \bar{\kappa}_0<0.\end{cases}$$

命题 3-1 中刻画了两个补贴政策之间复杂的关系。同时考虑这两项政策措施是非常必要的，孤立地考虑市场一侧而不考虑另一侧的补贴政策可能导致结果与最优策略有较大差距。同时，该结果中也刻画了当其中一个补贴最优为零的情况，此时政府的混合政策退化为单边补贴。接下来的推论中将更加详细描述此情况。

**推论 3-1：** 从命题 3-1 中的 TSSP 问题最优解中，可以得到如下推论：

（a）存在阈值 $r_\kappa$、$c_\kappa$、$p_\kappa$、$\bar{\theta}_\kappa$ 使得给予投资者的最优补贴为零，当且仅当下列条件之一满足：（1）$r < r_\kappa$，（2）$c > c_\kappa$，（3）$p > p_\kappa$，（4）$\widehat{\theta} < \widehat{\theta}_\kappa$；

（b）存在阈值 $r_s$、$c_s$、$p_s$、$\widehat{\theta}_s$ 使得给予消费者的最优补贴为零，当且仅当下列条件之一满足：（1）$r > r_s$，（2）$c < c_s$，（3）$p < p_s$，（4）$\widehat{\theta} > \widehat{\theta}_s$。

该推论说明了在哪些条件下最佳政策不会对市场双方都提供补贴。比较推论 3-1 的（a）和（b）会得出一些有价值的见解。在政府应仅采取直接措施（即补贴电动汽车消费者）与间接措施（即补贴基础设施投资者）以加快电动汽车采用过程的情况之间，形成了鲜明的对比。

推论 3-1 的最令人惊讶的结论可能来自市场两侧进入壁垒成本参数的影响，分别是电动汽车的零售价格和充电桩的初始投资费用。根据推论，这两个参数在驱动最佳政策的结构中起着相似的作用（即单边与两边）。这样的结论与直觉相反，因为在单方面政策被证明是最优的情况下，人们可能会期望政府的激励措施始终以进入成本较高的一侧为目标。推论 3-1 显示，情况并非如此。特别是，当两种成本中的任何一种超过其相应的阈值时，充电桩投资者都将被排除在政府支持之外，而仅激励消费者的单方面政策将成为最佳选择。另外，当这些参数中的任何一个都足够低时，就停止了对电动汽车消费者的直接补贴，并且通过促进基础设施建设可以更好地使用政府资源。

里程焦虑相关参数对最优补贴政策结构的影响更为直观。当里程焦虑参数足够低时，充电桩无法为消费者创造足够强大的网络效应。结果，向投资者提供经济激励措施不如直接补贴潜在的购买价格那样有利。但是，当网络效应变得足够强大时，情况恰恰相反，在这种情况下，电动汽车消费者不应获得任何激励。关于消费者的环境偏好也有类似的论点。

也就是说，当消费者的环保意识足够低时，决策者应该将其所有资源分配给购买补贴，从而直接提高消费者对电动汽车的效用。另一方面，当消费者对电动汽车的估值超过阈值时，停止给予消费者补贴并完全专注于充电基础设施将成为最佳选择。

在确定了最佳补贴以及电动汽车和充电桩的数量后，接下来研究模型参数变化时这些结果的表现。对于本书的其余部分，假定 $m_0 = 0$ 来简化敏感性分析。由于最优解在 $m_0$ 处是连续的，因此对于 $m_0 > 0$ 足够小的邻域，敏感性分析结果仍然成立。接下来首先分析电动汽车价格和潜在消费者人口规模的影响。

## 3.5　政 策 建 议

命题 3-1 刻画了给定科技水平和市场特点下的均衡情况时的最优补贴政策，也为研究外部环境变化时政府政策的变化奠定了基础。本小节研究在双补贴政策为最优策略的情况下，不同外部环境参数对最优补贴策略的影响。

### 3.5.1　技术成本降低

随着电动汽车在全球范围内的普及及电动汽车市场规模的不断扩大，电动汽车制造商的研究与开发实现了技术突破并降低了成本。例如，在过去的几年中，美国电动汽车的价格与化石燃料的价格相比呈下降趋势。[①] 充电桩的建设成本也呈现了类似的趋势，由于技术进步，该成本一直在下降（Nicholas，2019）。因此，了解这种趋势将如何影响政府补贴机制是非常重要的话题。命题 3-2 解决了这个问题。

**命题 3-2：** 技术成本降低对最优补贴政策及市场均衡的影响如下所示：

（1）均衡市场下的 $\lambda^*$ 和 $m^*$ 关于 $p$ 和 $c$ 都是递减的；

（2）消费补贴 $s^*$ 关于 $p$ 和 $c$ 递增；

（3）投资补贴 $\kappa^*$ 关于 $p$ 递减，并且关于 $c$ 是单峰函数。

① https://www.caranddriver.com/research/a31544842/how-much-is-an-electric-car/

命题 3-2 的第（1）部分表明，由于技术进步而在市场两边发生的任何成本降低都会对电动汽车的推广产生积极影响，因为这会增加电动汽车和充电桩的最终数量。较低的电动汽车价格或较低的充电桩投资成本分别吸引了更多的买家和投资者，从而促进了双方的增长。

更有趣的是，命题 3-2 的第（2）部分、第（3）部分表明，技术成本的降低以两种完全不同的方式影响了这两种补贴工具。随着电动汽车的零售价格或充电桩建设成本的下降，政府应减少对电动汽车消费者的补贴，这是可以预期的，因为较低的价格可使电动汽车在市场上更具竞争力，并在不需要太多政府支持的情况下促进其普及。而较低的建设成本可加快充电桩的投资建设，提高电动汽车的吸引力，并减少政策干预的需求。

虽然从直观上想，投资补贴 $\kappa^*$ 可能也会有类似的结论，然而通过命题 3-2 的第（3）部分却发现相反的策略可能更优。也就是说，当充电桩的建造投资成本降低的时候，给予投资者的补贴应该提高。这个结果背后的原因主要来自充电桩运营商的完全竞争。当电动汽车的价格越来越低的时候，使用电动汽车的人群数量也在增长，因此充电桩群体就可以获得更高的利润。假设此时的投资补贴 $\kappa^*$ 维持不变，将会有更多投资者进入充电桩市场。此时，政府如果增加 $\kappa^*$，充电桩投资者数量的增长率会更高。换句话说，当电动汽车的人口规模增加时，每增加一美元的投资补贴就将带来更多的电动汽车推广。因此，由于更好的电动汽车推广效果，政府将愿意提供更为慷慨的基础设施支持。

除此之外，充电桩补贴与充电桩成本之间的关系呈现了倒 U 形的非单调关系，这是由两股对抗的力量导致的。首先，更高的成本驱使政府提高其对投资者的补贴以降低进入壁垒；其次，由于上文提及的边际效应，当电动汽车数量比较少时，更高的补贴投入并不能有很高的收益。因此，当充电桩成本较高的时候，电动汽车和充电桩的数量比较少，第二项的作用更强，因此 $\kappa^*$ 上升而 $c$ 下降。然而当充电桩成本降到一定阈值以下的时候，第二项作用减弱，第一股力量占据主导地位，此时政府应该提高充电桩投资相关补贴。

因此，根据所处的科技发展水平及其对市场条件的影响，两种补贴既可能呈现替代作用，也可能呈现互补作用。展开来说，当电动汽车的价格和充电桩的成本比较适中的时候，$s^*$ 和 $\kappa^*$ 呈替代作用，朝着相反的方向变化；

而当充电桩成本持续下降的时候，$s^*$ 和 $\kappa^*$ 同方向变化，呈互补作用。

除此之外，命题 3-2 隐含了一个现实中经常被忽略的政策建议，对于只提供消费者补贴的国家和地区尤为重要。当电动汽车市场仍处于起步阶段并处于技术发展的早期阶段时（世界上大多数国家和地区都是这种情况），如果技术成本有所降低，那么应减少对电动汽车消费者的补贴，而提高对充电桩投资者的补贴。就是说，当电动汽车市场仍处于不成熟状态且远未具有竞争力时，电动汽车补贴应下降，而基础设施补贴实际上应随着时间的推移而增加。仅当充电桩投资成本充分下降时，政府才应降低对基础设施的补贴。

### 3.5.2　网络效应

电动汽车的吸引力在很大程度上取决于公共充电站的可用性。随着充电基础设施的铺开和更多充电桩的投入使用，消费者将能够更轻松地访问充电设施，并驾驶电动汽车行驶更远的距离。这种积极的网络效应减轻了里程焦虑，增强了消费者的电动汽车驾驶体验。但是由于驾驶习惯、城市特征和出行距离等因素，这种影响的程度可能因不同的消费者群体而异。因此政府在设计激励政策和调整补贴时需要考虑网络效应的强度。相应地，在本模型中，网络效应通过参数 $r$ 表示，该参数反映了随着站点数量的增加，消费者的里程焦虑缓解的速率。

**命题 3-3：** 更高的 $r$ 带来更低的消费者补贴 $s^*$ 和更高的充电桩投资补贴。与此同时，更高的 $r$ 可以使政府总体更优，也就是带来更高的 $\Pi^*$、$\lambda^*$ 和 $m^*$。

该结果表明，在里程焦虑更强或者消费者对充电桩可用性更加敏感的地区，政府应增加对基础设施发展的补贴，并减少对电动汽车消费者的补贴。较高的 $r$ 会增加电动汽车对潜在消费者的吸引力，并引起更多的电动汽车购买。尽管这样做可能减少最优消费者补贴 $s^*$，但同时也增加了充电桩投资者的总收入，并激励了更多的投资者进入市场。直觉上，随着网络效应变得更强，$\kappa^*$ 与 $s^*$ 都呈下降趋势。但是，命题 3-3 却显示了相反的结果，随着 $r$ 增长，对投资者的补贴应增加。此发现更适用于平均行程较长且长途通勤是人们日常生活的一部分的地区（例如，南加州、得克萨斯州）。在这种情况下，政策制定者应该利用更强大的网络效应，为

充电桩提供更多慷慨的补贴，因为该策略最终将推动电动汽车的普及，并造福于政府。

值得补充的是，参数 $r$ 还可以代表电动汽车电池技术的改进程度，它决定了电动汽车无须充电即可行驶的总距离。随着技术的发展和电动汽车电池能够容纳更多的能量，消费者对充电桩网络密度的敏感度将下降，网络效应减弱。此时，命题 3-3 建议将政策重点从充电桩建设转移，并加强电动汽车购买补贴。该建议与传统观点背道而驰，因为很多政府在电池容量和使用寿命提高的同时，通常会降低对电动汽车消费者的补贴。

### 3.5.3 环保意识

不同的消费者和政策制定者对环境的偏好和应对气候变化的态度差异很大。这反映了特定地区对环境保护态度的程度，也体现了消费者的环境意识和政府对绿色技术的重视程度。例如，在美国，居民的环境保护意识在不同州之间存在高度异质性。同样，对于空气污染和雾霾严重的地区，政府可能对诸如电动汽车等环保产品有更高的评价。这种差异启发我们考虑如何将不同利益相关者的环境偏好纳入政府对电动汽车的激励政策中。

接下来的两个命题中，通过计算模型中的相关参数来回答上述问题。参数 $\beta$ 表示电动汽车给社会的环境和健康带来边际提升。参数 $\widehat{\theta}$ 表示了顾客使用有利环境的交通工具给消费者带来的效用提升。当 $\widehat{\theta}$ 增高的时候，消费者平均环保意识有所提升，其环保态度的方差也有所提升。

**命题 3-4：** 最优充电设施补贴 $\kappa^*$ 关于 $\beta$ 单调递增。存在一个网络效应的阈值 $\widehat{r}$，当且仅当 $r \leqslant \widehat{r}$ 时，最优消费补贴 $s^*$ 关于 $\beta$ 单调递增。

该结果表明，随着电动汽车对社会的利益增加，政府应始终加大对充电桩的支持力度。但是，对于支持电动汽车消费者而言，同样的结论不一定正确。尤其是当网络效应足够强时，较高的电动汽车社会效益转化为较低的电动汽车购买补贴。在这种情况下，向充电桩提供更慷慨的补贴将比补贴电动汽车消费者更有效地提高电动汽车的推广。因此，当 $\beta$ 增大的时候，政府最佳的做法是减少对消费者的补贴。这意味着政府应该根据电动汽车对健康和环境的影响，仔细调整其补贴计划。

**命题 3-5：** 存在关于充电桩建造成本的阈值 $\widehat{c}$，满足：

（1）如果 $c > \widehat{c}$，那么随着 $\widehat{\theta}$ 的增大，$s^*$ 降低而 $\kappa^*$ 上升；

（2）如果 $c \leqslant \widehat{c}$，那么 $s^*$ 是关于 $\widehat{\theta}$ 的单峰函数，而 $\kappa^*$ 随着 $\widehat{\theta}$ 上升而下降。

命题 3-5 中展示了消费者的环保意识依赖于充电桩的投资成本可能会带来不同的补贴效应。当 $\widehat{\theta}$ 的值更高时，政府可以从两方面利用该效应。一方面，当人群的环保意识更强，顾客购买电动汽车的效用更高，因此政府可以减少对电动汽车消费者的补贴，将支出更多地转移到充电桩投资者一方来更好地促进推广。在此情况下，政府的总支出实际上上升了，但是支出的上升会被更高的收益所弥补。

另一方面，政府也可以通过更高的 $\widehat{\theta}$ 来节省开支。此时，政府可以降低充电桩补贴，因为节省的支出要比失去的那部分电动汽车带来的环保收益更高。通过命题 3-5 可知，当充电桩的成本比较高的时候，适用于第一种情况，而第二种方案更适合成本比较低的时候。因此，消费者的环保意识既有可能提高两种补贴水平，也可能降低两种补贴水平。

消费者的环保意识变化对最优决策的影响取决于充电桩的成本。当投资成本高昂时，消费者的环保意识更高，政府就可以将其补贴从消费者转移到投资者，从而促进电动汽车的普及和充电桩数量的增加。当投资成本较低时，当消费者在一定程度上变得更加环保时，政府就会增加对消费者的支持。

## 3.6　模 型 验 证

尽管电动汽车的推广效果不及预期，但是中国仍然是世界上电动汽车保有量最大的国家，而深圳则是中国电动汽车保有量最大的城市。[①] 深圳政府是最早开始采取措施推广电动汽车的城市之一，于 2009 年开始了补贴政策。当时市场上只有极少数的电动汽车车主，而充电桩的数量更是屈指可数。[②]该补贴政策持续到了 2014 年，2014 年之后深圳对小汽车的上牌做了限制，每年只发布 10 万辆传统汽车牌照，而对新能源汽车没有限制。

① https://www.tyncar.com/schq/0327-8298.html

② http://www.sz.gov.cn/cn/xxgk/zfxxgj/zwdt/201007/t20100707_5292871.htm

本部分搜集了多个来源的数据来验证提出的模型，[①] 相关的参数估计如表 3-2 所示。将消费者外部选择归一化，$u_0 = 0$。深圳有几种销量比较高的电动汽车车型，取其价格平均值作为模型中的 $p$。充电桩的建造成本来源于哥伦比亚大学的研究报告（Anders & David，2019）。虽然一些参数可以从现有数据中很容易地得到（比如补贴的金额、电动汽车价格、充电桩成本等），但是其他的一些参数很难得到。例如，用每年所有汽车的总销量（包括传统汽车和电动汽车）作为对市场规模 $\Lambda$ 的估计。网络效应的相关参数及政府和居民对环保的重视程度的相关参数都很难估计。

**表 3-2 参数的估算值**

| 变量 | 值 | 数据来源 |
|---|---|---|
| $p$ | 250000 | 电动汽车价格 |
| $\phi$ | 1800 | 平均充电费用 |
| $c$ | 40000，500000，60000 | 充电桩建设成本 |
| $\Lambda$ | 487354 | 潜在市场规模 |
| $r_0$ | 96444.96 | 里程焦虑上限 |
| $r$ | 70.16 | 网络效应 |
| $\hat{\theta}$ | 366583.65 | 消费者环保意识 |

为了估计这几个参数，本部分采取了以下方式。首先假设深圳政府使用该模型进行决策。获取了 2014—2018 年的如下数据：

（1）年初电动汽车的初始数量 $\lambda_0$，充电桩的初始数量 $m_0$；

（2）消费者补贴 $s$，充电桩投资者补贴 $\kappa$；

（3）年终电动汽车数量 $\lambda$，充电桩数量 $m$。

使用（1）和（2）中的数据，可以代入未知参数计算本书提出的模型，得到带有这几个参数的年终电动汽车数量和充电桩数量。接下来，比较带有参数的估计和实际的数量，并通过最小化估计误差的优化模型求解相应的参数。

网络效应 $r$ 是本模型中最重要的参数之一，能够反映消费者里程焦虑的情况，因此本书尝试了不同的参数范围来观察其对模型的影响。假设 $r = \{50，60，\cdots，200\}$，并且满足假设 3-1 中的不等式。模型中最难以

① http://www.xinhuanet.com/fortune/2018-07/22/c_1123160496.htm，https://tech.sina.com.cn/i/2018-09-16/doc-ifxeuwwr4866727.shtml，http://sz.people.com.cn/n/2014/0114/c202846-20379736.html

估计的参数为衡量政府估计的单位电动汽车的社会价值的参数 $\beta$。因此应该允许 $\beta$ 可以选择广泛的参数范围 $\beta \in \{200000,\ 300000,\ 400000,\ 500000\}$。为了检验本模型的稳健性，本小节考虑了以上 $\beta$、$c$ 和 $r$ 的不同参数组合，分别进行了实验。

对于模型参数的每种可能组合，考虑两种可能的情况：实施电动汽车补贴和充电桩投资补贴的混合政策，以及实施仅补贴电动汽车的单方面政策。本部分找到了两种方案的最佳补贴政策，并比较了它们相应的结果。具体来说，本书计算了两种补贴策略下，电动汽车推广数量、充电桩以及政府目标函数的比例。在所有的实验中，混合政策下的充电桩的平均数量是仅提供消费补贴的平均数量的 11.26 倍，其中最大的倍数达到了 33.96。混合策略下的电动汽车数量平均增长了 6.82%，最高可达到 30%。政府的目标函数平均增加了 5.68%，最高增加了 26%。

图 3-2 和图 3-3 展示了此数值实验的两个例子。其中图 3-2 展示了

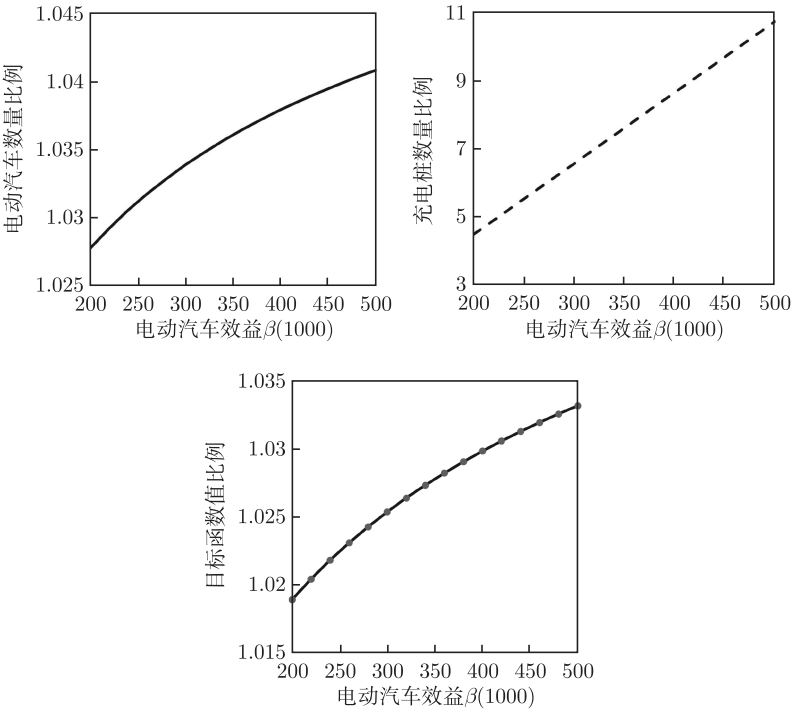



**图 3-2　混合补贴政策与单边补贴政策的对比 ($c = 50000$，$r = 100$)**

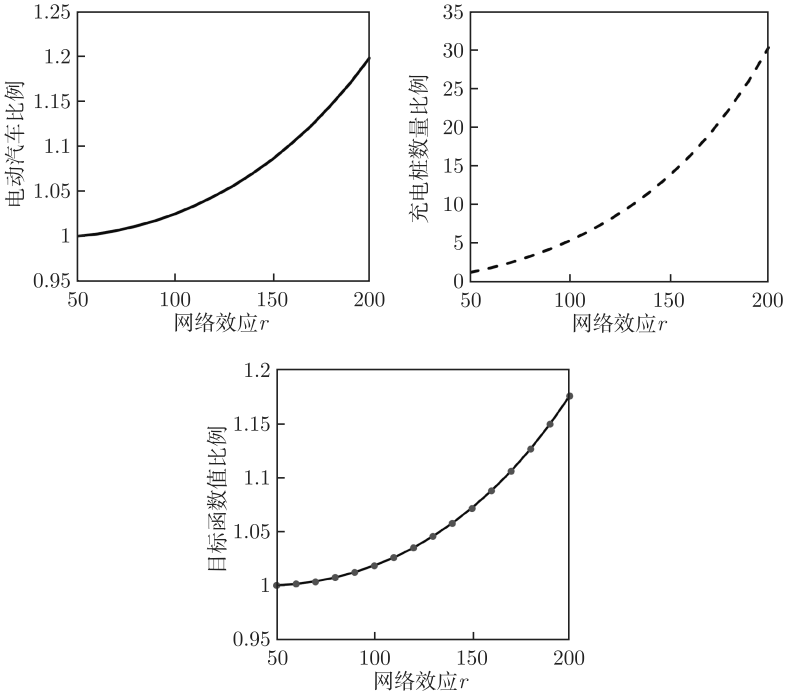



**图 3-3 混合补贴政策与单边补贴政策的对比 ($\beta = 300000$，$c = 60000$)**

电动汽车社会收益（参数 $\beta$）变化对三个比例的影响。可以发现，当同时补贴电动汽车消费者和充电桩投资者时，均衡下的充电桩数量是只对消费者进行补贴时的 4~10 倍，同时，这也带来了电动汽车数量 2.7~4.1 个百分点的提升，也给政府的目标函数值带来了 1.8~3.3 个百分点的提升。

图 3-3 阐述了当基础设施成本为 60000 时，网络效应参数 $r$ 对几个比例的影响。更高的网络效应参数 $r$ 代表着更高的里程焦虑，因此提高充电桩的数量能够更有效地提高消费者购买电动汽车的效用。当消费者的里程焦虑较弱时（$r$ 比较低），政府的最优策略为只提供消费者补贴。反过来，当消费者的里程焦虑较强时，提供混合补贴政策会更好。实际上，随着 $r$ 的增大，政府会减少对消费者的补贴而增加对充电桩投资者的补贴。这样对补贴政策的调整可能导致巨大的均衡下充电桩的数量变化。此时，电动汽车数量能够增长 20 个百分点，而政府目标函数值能够增长 17 个百分点。这个例子也从另一方面验证了政府应该仔细考虑网络效应的强

度，设计更为合理的补贴策略。

## 3.7 结　　论

气候变化为世界带来了严重的影响，倒逼决策制定者不得不付出努力来推广环保技术。电动汽车能够有效地减少温室气体的排放并保护环境。然而电动汽车的推广效果却尚不令人满意，其中两个至关重要的因素是电动汽车的高价格及缺少公共充电桩导致的里程焦虑问题。为了克服这样的问题，很多政府引入了不同的激励方案来增加电动汽车的吸引力，促进电动汽车的推广。不同政府的策略各有不同，有的政府只对电动汽车进行补贴，使消费者有力承担其价格；有的政府采取了混合的策略，一方面为消费者提供补贴，另一方面也对充电桩的投资者进行补贴，吸引更多的充电设施建设。

不同政府的不同做法启发本部分对政府的最优补贴政策的研究。本模型考虑了充电桩投资者之间的竞争及电动汽车市场的双边效应。更多的电动汽车会给充电桩投资者带来更多的利润，而更多的充电桩也会提升电动汽车的效用，两者形成了正向的网络效应。在这样的市场结构下，计算出了政府的最优补贴政策，并给出了政府应采取混合策略而非单边补贴策略的条件。可以发现，当网络效应比较弱，充电桩成本或电动汽车成本比较高，消费者的环保意识并不强的时候，采取单边的补贴政策是更好的选择。这对决策制定者的补贴策略选择具有重要的指导意义。

除此之外，本书还考虑了不同市场条件及科技变化时的政策调整方案。具体来说，分别考虑了科技成本降低、网络效应变化、政府及消费者环保意识的变化对最优补贴政策的影响。结果显示，这两种补贴政策取决于科技提升，既有可能呈现互补性也可能呈现替代性。具体来说，当电动汽车市场处于初期阶段时，当电动汽车的价格逐渐增长，政府应该逐渐降低对消费者的补贴，而增加对投资者的补贴。此条建议同样适用于网络效应增强、顾客的里程焦虑增强之时。当政府更加在乎环保的时候，政府应该给予投资者更多补贴，但是消费者方面的补贴取决于网络效应的强度。如果网络效应比较强，那么政府应该减少对消费者的补贴，反而通过增加投资者补贴间接促进电动汽车推广。另外，消费者的环保意识对最优补贴

政策的影响同时也受到充电桩成本的影响。

最后采集了深圳的数据验证了本书的模型。可以发现混合策略对于深圳政府是最优策略，相比单边策略平均能够提升 6.82%的电动汽车数量和 5.68%的政府效用，最高甚至分别可达到 30%和 26%的提升。另外，混合策略下的均衡市场的充电桩数量平均达到了单边市场数量的 11.26 倍，在一些极端场景下甚至可以高达 34 倍。这些结果表明，采用混合策略相比单边策略可能带来更大的潜在收益，这也是当前很多国家都采用了混合策略的原因。

此部分仍有可改进的空间。首先，由于分析的复杂性，本部分没有考虑消费者的战略等待或政府的动态补贴优化，而将分析集中在讨论网络效应和两种补贴之间的互补性上。同时，将消费者、投资者和政府的战略行为纳入多个时期的模型将是很有趣的，但是可能需要简化假设使得此动态博弈问题可解。其次，研究电动汽车供应链中政府补贴与制造商决策之间的相互作用会是很有趣的方向。这些决策包括对电池技术的创新和研发投资、供应商的市场规模决策，以及制造商对电动汽车的定价。最后，本部分的理论框架适合进行全面的实证研究，可以利用有关消费者和投资者决策过程的微观基础的更详细的数据，验证分析结果的稳健性。希望此部分的工作能为电动汽车运营管理领域提供更多的研究参考。

# 第 4 章　数据中心供应链的网络设计与服务资源分配

本部分以数据中心供应链为例，研究了作为互联网相关服务和云计算的物理基础架构的数据中心的网络设计与资源供给问题。数据中心的建造者需要控制成本，同时提高服务质量以获得竞争优势。本部分为数据中心网络设计与基础设施资源分配问题构建了数学模型。模型通过优化数据中心的位置、需求分配和资源供应决策，达到总运营成本和服务延迟损失的最小化。模型的拓展中也考虑了诸如延迟、功耗、多种资源、配置限制以及相互依赖的需求等问题。针对服务延迟问题，本部分创新地采用排队模型进行估计，并对延迟做了变形，将问题转化为二阶锥优化问题。为了提高大规模问题的计算效率，本部分创新地开发了两种基于拉格朗日松弛的算法，一方面设计了有上界表现的启发式算法，另一方面也分析了问题的结构特性来生成加强割平面。数值研究表明，本部分提出的解决方案优于最新的商业软件。基于现实世界的数据集，应用本部分的模型选择的数据中心与主流云计算服务提供商选择的数据中心不谋而合。最后通过对参数进行敏感性分析的数值实验，得到了丰富的数据中心网络设计和服务资源分配的管理启示。

## 4.1　引　　言

在过去的 20 年中，互联网相关的服务和云计算的需求激增，对托管在数据中心的基础设施提出了更高的要求。2017 年，北美地区主要服务提供商对数据中心的投资总额超过 200 亿美元。[①]在 2018 年第三季度的

① https://www.cbre.us/research-and-reports/US-Data-Center-Trends-H2-2017

收益报告中，Alphabet Inc.（谷歌的母公司）报告称，该公司前九个月的资本支出同比增长了一倍以上，达到 186 亿美元（相比之下，运营支出增长了 25%），而大部分支出用于数据中心建设。[①]

云计算和其他相关服务市场机遇吸引了包括谷歌、亚马逊、微软和脸书在内的互联网巨头，这些巨头们已经开始逐渐开放其数据中心，向其他公司提供云计算服务，服务对象包括小型初创公司，也包括类似优步、网飞等估值数十亿的独角兽公司。激烈的竞争迫使基础设施提供商提供质量更高、成本效益更优的服务。为了实现这一目标，基础设施提供商无疑需要更好的数据中心网络设计与基础设施资源的分配策略。

互联网和云计算行业的数据中心网络设计与制造业和零售行业的供应链和物流网络设计有着异曲同工之处，同样占据着重要地位。但是，过去学术界更多关注传统的供应链网络。尽管已经有很多复杂模型来帮助公司优化其供应链和物流网络，但是人们对最佳数据中心网络的理解仍然非常有限。因此，本书旨在为数据中心网络设计问题提供决策工具，优化数据中心的位置、服务资源的分配以及供应（即容量规划，分配不同的资源来满足计算和存储资源的需求）等。对于拥有大量基础设施投资和运营成本居高不下，同时试图努力保持较高的客户服务水平的公司来说，这些决策至关重要。尽管数据中心网络和供应链网络设计具有许多共同的特征，例如，两者都具有大型且昂贵的设施，并且运营成本对设施的位置以及设施与需求点之间的距离敏感，但数据中心网络的设计所具备的很多特点也更具挑战性，下文将对此进行详细说明。

首先，与传统的配送中心不同，数据中心消耗大量的电。例如，谷歌在 2011 年披露其数据中心占全球用电量的 0.01%。[②] 因此，与位置有关的电力成本对于数据中心网络而言意义重大。其次，在数据中心网络中，存在共享有限容量的多种资源，例如计算存储资源等。此外，对于各种资源类型，不同的需求点可能具有不同的要求。因此，数据中心网络中的资源供应和需求分配问题比供应链网络的挑战性更大。再次，在数据中心网络中可能引起两种类型的延迟——数据中心内的延迟（即终端主机等

① https://abc.xyz/investor/pdf/20181025_alphabet_10Q.pdf

② https://www.theguardian.com/environment/2016/jul/20/google-ai-cut-data-centre-energy-use-15-per-cent

待时间）和需求点与其分配的数据中心之间的延迟（即网络延迟等待时间）。终端主机的等待时间与资源供应和需求分配的关系是非线性的，并且与网络传输延迟时间同等重要。相比之下，对于供应链网络，人们通常将注意力集中在传输时间（类似于网络延迟）上，而忽略了分发中心内的处理时间（类似于终端主机延迟）。

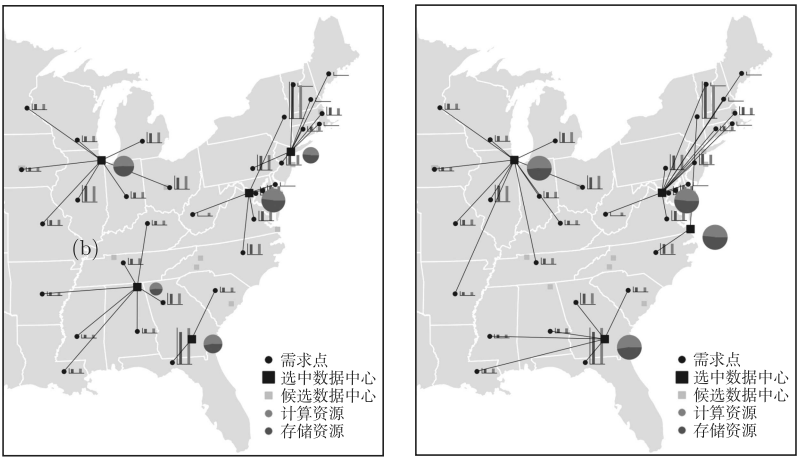
除上述挑战外，数据中心网络的一些不同于传统供应链的特征使得设计问题进一步复杂化。其中日益重要的一个特征是需求点之间由于同步或冗余备份所导致的相互依赖性。在这种情况下，在一个数据中心处理的任务将以一定的概率需要启动另一个数据中心的另一个任务，从而增加了数据中心之间的额外工作量和流量。除此之外，功耗与平均工作负载之间、架构延迟与平均流量之间通常存在非线性相关性，这也对模型的求解提出了挑战。

忽略上述的数据中心网络设计的特征有可能导致不合理的决策。因此，通过优化数据中心选址、需求分配、资源分配等决策，综合考虑最小化总运营成本和延迟成本，以及其他数据中心网络的特点，提出了数据中心网络设计模型。值得注意的是，整合优化战略层和运营层决策在过去的文献中屡见不鲜。Daskin et al.（2002）和 Shen et al.（2003）首先提出了整合考虑选址和库存的模型，模型中考虑了非线性的安全库存水平，并给出了最优的选址决策和库存订货决策。类似地，文献中有很多其他类型的整合模型，比如 Geunes et al.（2011）研究了供应链管理的整合模型，Taaffe et al.（2008）研究了营销中的整合模型，Geunes & Pardalos（2003）综述了供应链以及金融工程中涉及的相关整合模型。

考虑到问题的复杂性，当前数据中心网络设计诉诸简化模型，例如使用分层优化的方法。具体而言，将整合模型分解为多个阶段，先进行运营层面的优化，再进行战略层面的优化，但是这样做的代价是可能会得到次优的解。例如，云基础设施提供商的新数据中心的位置是根据例如当地电价、温度（影响冷却需求）以及对可再生能源的使用等因素决定的，但是未考虑其他例如需求分配、资源供应和服务质量等因素。在确定了数据中心的位置之后，再根据需求预测估算对 CPU 和存储等资源的需求，并将其分配到不同的数据中心。Barroso et al.（2013）总结了当前数据中心选址以及网络设计做法的详细信息。相比之下，本书使用了高度集成的模

型，与大多数现有模型相比更适合数据中心网络设计。

接下来，可以使用一个简单的数值例子来展示整合模型的优势。假设需要从一些候选位置中选择一些数据中心来服务需求，利用整合模型可以同时最小化总运营成本和延迟。同时，考虑以一个分层规划的模型作为基准。首先构建一个简化的不考虑延迟的情况下的模型，算出数据中心的选址和网络布局，接着再优化资源配置决策来最小化延迟以及能源消耗成本。分层优化的具体模型可参见附录 C.3.1。图 4-1 对比了分层模型和本部分所提出的模型的选址、需求分配和资源供给的结果。不同需求点对资源的需求用图中需求点旁的柱状图表示。图中数据中心旁边的饼状图描述了每个数据中心所提供的资源比例，其中饼状图的大小表示了总的能源消耗。基准模型 [图 4-1(a)] 不仅相比最优模型 [图 4-1(b)] 额外多使用了一个数据中心，而且造成了双倍的终端延迟成本，总成本提高 33%。这些结果表明，整合的数据中心网络设计模型相比分层优化模型能够更加有效地平衡成本和服务水平，在竞争激烈的市场中有很大的潜力。最后，本部分也通过数值实验验证了模型在一些参数变化时，比如在需求增长的情况下，也能保证很好的鲁棒性。接下来将简要对相关文献进行梳理和回顾，并通过对比现有研究，总结本部分的贡献。




（a）基准分层模型　　（b）整合模型

**图 4-1　两模型的数据中心网络设计比较（见文前彩图）**

注：整合模型节省了超过 1/2 的终端延迟成本和超过 1/4 的总成本。

## 4.2　相关文献及贡献

近些年来，数据中心网络设计的问题得到了学界和业界的重视。Iyoob et al.（2013）综述了云计算和数据中心运营管理问题的挑战与机遇，提出了服务提供商视角下运营问题的层级。其中数据中心的选址、容量规划（即资源供给）是最重要的问题之一。Greenberg et al.（2008）也将数据中心的位置选择和建筑面积（资源供给水平）作为数据中心管理中最重要、最具挑战的问题之一。这个问题也吸引了计算机领域的关注，比如 Larumbe & Sansò（2012）、Larumbe & Sansò（2013）及其引用文献，但是大部分研究只考虑了数据中心的部分运营特点，并且依赖于计算复杂度较高的算法来找到局部最优解。

目前学术界主要研究云计算服务的资源管理问题。其问题的关键在于通过优化服务器负载和资源分配，在同时满足服务水平协议的情况下提高运营效率 [例如 Verma et al.（2008）和 Goudarzi et al.（2012）及其参考文献]，更贴近于本书的第三部分的研究内容。然而，本部分的工作关注整个数据中心的层级而不是服务器层级，而且考虑了其他的一些例如主机托管等问题。数据中心设计模型还引入了对传输和终端延迟的建模、电量与需求间的非线性关系，放松了假设以获得更好的模型准确度。

在运筹与管理科学领域的文献中，本部分的模型属于考虑拥堵的不可移动的服务资源选址问题。这类问题与传统选址问题的主要不同之处在于，模型一般假设设施需要用有限资源来满足服务时间随机性的需求。Berman & Krass（2015）对该领域的文献进行了详尽的回顾和分类。在此文章的分类下，本部分的模型属于顾客不会主动选择服务设施地点，而且需求速率固定这一类问题。在这方面的研究中，本书的模型结合了顾客与数据中心网络所产生的成本，并试图从中找到平衡。本工作通过提出一种新颖的模型来适应数据中心网络设计带来的上述挑战，例如将电力成本作为利用率的考虑、资源供应、架构和最终主机延迟以及它们对资源供应的依赖关系，以及需求足迹之间的相互依赖关系等，从而为对拥挤和服务资源不流动的设施选址问题的快速增长的研究做出了贡献。本书还通过探索模型的结构特性提供有效的解决方案，为整数规划相关领域做出了贡献。

本部分的模型尤其与 Berman & Krass（2015）中提到的具有平衡目标的模型紧密相关，其中包括 Wang et al.（2004）、Elhedhli（2006b）、Aboolian et al.（2008）、Castillo et al. （2009）、Zhang et al.（2009、2010）、Abouee-Mehrizi et al.（2011）、Paraskevopoulos et al. （2016）等。展开来说，Zhang et al.（2009）在医疗设施网络中考虑了服务的拥堵情况，通过优化选址决策最大化对患者的服务水平。其中顾客等待时间使用了 $M/M/1$ 模型进行刻画。Zhang et al.（2010）在上述模型的基础上进行了拓展，考虑了资源容量的决策。他们使用 $M/M/s$ 排队模型来估计顾客等待时间，并将服务台数量 $s$ 作为决策变量。

相比于该领域的其他文献，本书模型有如下特点。首先，采用具有处理器共享调度准则（processor sharing）的排队网络模型，该模型更适合与互联网相关的服务和云计算的应用，能够对多种资源类型和相互依赖的需求点进行建模。其次，除了终端延迟（即设施中的拥塞）之外，本部分还考虑了网络拥塞和其他关键因素，以实现对数据中心运营的更准确建模。将这些功能整合到现有模型中可能会给建模和计算带来不小的挑战。再次，本书提供有效的混合整数二阶锥规划（MISOCP）模型，并根据问题的结构性质开发量身定制的求解方法，而大多数现有研究依赖于更通用的方法，例如线性化（Aboolian et al.，2007）、列生成（Elhedhli，2006a）或启发式算法（Zhang et al.，2009、2010）来解决非线性优化问题。

最近，MISOCP 已被广泛应用于解决许多传统或新兴的运营管理问题，包括具有生产决策的鲁棒选址问题（Baron et al.，2011）、集成供应链网络设计（Atamtürk et al.，2012）、电动汽车电池交换基础设施规划（Mak et al.，2013）、预约时间表规划（Kong et al.，2013；Mak et al.，2015；Kong et al.，2020）、风力发电储能和输电联合规划（Qi et al.，2015）、共享出行系统设计（He et al.，2017、2019）和选品优化（Sen et al.，2019）等。这些模型中的大多数集中于对需求方差/协方差或概率约束进行建模，而本书将应用领域扩展到了随机服务系统的拥塞问题中。此外，本书没有直接使用商业软件解决 MISOCP 问题，而是开发了有效的拉格朗日松弛算法来加快求解速度。

本部分的主要贡献总结如下：

（1）本部分提出了一种新颖的数据中心网络设计模型（见 4.3 节），

并附加了三个模型扩展。同时考虑到对实际数据中心网络设计有重要影响的因素，包括多种资源类型、异质需求组合、网络与终端延迟、服务器配置受限引起的资源比率约束、主机代管、相互依赖的占用空间、非线性的功耗和网络拥塞等。通过文献对比可知，这是第一个综合考虑这些因素的模型，该模型适合实际的数据中心网络设计。

（2）本部分对提出的模型进行了一些变换，使原模型转化为 MISOCP 问题（4.4.1 节），并提供了拉格朗日松弛方法来解决变形后的问题（4.4.2 节），使用了加强割平面的方法（extended extremal polymatroid cuts）进一步提高了求解速度（4.4.3 节）。本部分的数值结果表明，使用这些算法可以明显优于当前的常用商业软件求解速度（4.5.4 节）。现实生活中，很多商业公司例如脸书 [①] 正在扩大其数据中心覆盖网络，甲骨文[②]也正在逐步规划建立自己的数据中心网络，因此本部分的模型和解法对这些数据中心提供商而言有很重要的价值。

（3）通过将现实世界的数据应用到本部分的模型中，得到了以下数据中心设计的启示（4.5.1 节）：（a）更高的单位传输延迟成本总是导致建设更多的数据中心，而更高的终端延迟成本可能会让需求分配到更少或更多的数据中心，具体取决于电源容量和资源比率限制；（b）对于需求量适中、电费成本低且距常规数据中心的距离较远的地区，最好分配到托管中心进行服务；（c）不同需求点之间的需求的高度相互依赖关系导致一些有很强相关性的数据中心距离较近。这些数值研究还证明了集成建模在数据中心网络设计中的重要性。

## 4.3　模　　型

本节将提出一种用于设计互联网相关服务或云计算公司的数据中心骨干网的优化模型。首先，制定确定数据中心位置、需求分配和资源配置的基础模型；其次，引入一个排队模型来估算终端主机的延迟成本；最后，提出几种模型扩展。

① https://datacenterfrontier.com/facebooks-accelerates-data-center-expansion/

② http://www.datacenterknowledge.com/archives/2016/01/13/linear-programming-helps-groupon-optimize-data-center-design

### 4.3.1 基础模型

假设需求来源于需求点的集合 $\mathcal{I}=\{1,\ \cdots,\ |\mathcal{I}|\}$。每个需求点 $i\in\mathcal{I}$ 可以被认为是处于同一地理区域内，对于各种服务（如邮件、社交网络、网络游戏等）有需求的顾客集合。假设这些需求被打包成数据包（作业）传送到数据中心进行处理，需求按照独立的泊松过程到达，需求点 $i$ 的到达率为 $d_i$。处理一个数据包（作业）需要一些资源，用 $\mathcal{K}=\{1,\ \cdots,\ |\mathcal{K}|\}$ 表示。其中每种资源处理一些特定的任务，比如计算机中央处理器（CPU）的资源用来处理计算类型的需求、硬盘资源用以满足存储读取需求、显卡资源用来处理机器学习相关需求等。为了便于说明，本书假设 $|\mathcal{K}|=2$，而这两种资源则分别是计算资源和存储资源，该模型可以很容易地拓展到有多种类型资源的情况中去。与此同时，假设数据包的大小可以表示其对资源的需求量，并假设平均的数据包大小已知。用 $u_{ik}$ 来表示需求点 $i$ 对资源 $k$ 的需求量。不同需求点对不同资源的需求比例可能会有很大的差别。例如，有的地区的用户可能观看视频更多，所以对存储资源的需求更大；而另外一些地区的用户可能更多使用搜索或者云计算等功能，所以对 CPU 计算资源的需求更大。

数据中心网络设计的问题需要考虑以下决策。从候选数据中心集合 $\mathcal{J}=\{1,\ \cdots,\ |\mathcal{J}|\}$ 中选出需要建设的数据中心。令 $\boldsymbol{x}=(x_1,\ \cdots,\ x_{|\mathcal{J}|})$，表示数据中心选址决策。其中，如果选择了数据中心 $j$，则 $x_j=1$，否则 $x_j=0$。每个顾客的服务需求会沿着数据中心网络被分配到指定的数据中心去。令 $\boldsymbol{y}_{\cdot j}=(y_{1j},\ \cdots,\ y_{|\mathcal{I}|j})$，表示数据中心 $j$ 的分配决策。其中，如果需求点 $i$ 的需求被分配到数据中心 $j$ 进行处理，那么 $y_{ij}=1$，否则 $y_{ij}=0$。除此之外，也需要优化每个数据中心内的资源供给。令 $\boldsymbol{z}_{j\cdot}=(z_{j1},\ \cdots,\ z_{j|\mathcal{K}|})$，表示资源供给决策。其中，$z_{jk}$ 表示数据中心 $j$ 中可以提供资源 $k$ 的总量。数据中心可以提供的总资源被其耗能上限所约束，而且由于实际中的服务器配置类型有限，不同资源之间的比例也有上限和下限的比例约束。令 $\bar{r}_{kl}$ 表示在同一个数据中心中资源 $k$ 与资源 $l$ 之间的最大比例。

本部分考虑最优化如下成本。在候选点 $j$ 建造数据中心会带来固定成本 $f_j$。类似建设数据中心这种巨额开支，投资者一般会采用贷款等方式，所以固定成本可以摊销到每年。数据中心 $j$ 的单位用电成本为 $c_j$，其

中用电成本不仅包括电费，还包括该数据中心的制冷以及其他维护所需成本。为了将服务质量也整合到模型的目标中，同时也需要考虑传输延迟和终端服务延迟成本。Shen et al.（2003）和 Berman & Krass（2015）等也采用了类似的整合方法。令 $t_{ij}$ 表示从需求点 $i$ 到数据中心 $j$ 的单位传输延迟成本，令 $L_j(\boldsymbol{y}_{\cdot j}, \boldsymbol{z}_{j\cdot})$ 表示在给定需求分配决策 $\boldsymbol{y}_{\cdot j}$ 和资源供给决策 $\boldsymbol{z}_{j\cdot}$ 时，数据中心 $j$ 导致的总终端延迟成本。

本章所使用的主要符号如下表 4-1 所示。

**表 4-1　本章的主要符号**

| 符号 | 解释 |
|---|---|
| $\mathcal{I}$ | 需求点的集合 |
| $\mathcal{J}$ | 候选数据中心集合 |
| $\mathcal{K}$ | 需要的资源类型集合 |
| $d_i$ | 需求点 $i$ 的需求到达率 |
| $u_{ik}$ | 需求点 $i$ 对资源 $k$ 的单位需求率 |
| $p_j$ | 数据中心 $j$ 的耗能上限 |
| $c_j$ | 数据中心 $j$ 的单位耗能成本 |
| $f_j$ | 数据中心 $j$ 的固定建造成本 |
| $t_{ij}$ | 从需求点 $i$ 到数据中心 $j$ 的单位传输延迟成本 |
| $t_{jj'}$ | 从候选数据中心 $j$ 到候选数据中心 $j'$ 的单位传输延迟成本 |
| $\tau_i$ | 需求点 $i$ 的单位终端延迟成本 |
| $w_k$ | 单位资源 $k$ 的峰值能源消耗 |
| $\alpha$ | 峰值能源消耗占比 |
| $\bar{r}_{kl}$ | 资源 $k$ 与资源 $l$ 在同一个数据中心的最大比例限制 |
| $x_j$ | 是否选择候选数据中心 $j$ |
| $y_{ij}$ | 需求点 $i$ 是否被分配到数据中心 $j$ |
| $z_{jk}$ | 数据中心 $j$ 提供的资源 $k$ 的总量 |

本部分的模型拓展也考虑了微型托管数据中心（也称为 Colo）。实际上，数据中心的服务提供者可以与当地数据仓库提供商签约托管，以服务特定区域的用户，从而避免巨大的资本支出并减少传输延迟（Greenberg et al.，2008）。因此，可以添加一组决策变量 $\{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{z}}\}$。在下文中，使用 $\tilde{\cdot}$ 来表示托管相关决策的变量。本书假设托管数据中心仅服务于本地需求，

并且可在所有需求点使用。因此，如果需求点 $i$ 被分配到了当地的托管中心，那么 $\tilde{x}_i = 1$。类似于数据中心的固定成本，托管数据中心的租用成本可以按年摊销或租金进行衡量，因此可以与其他目标函数中的成本综合考虑。值得注意的是，为了避免需求点 $i$ 的需求要么被托管中心服务，要么被数据中心服务的情况，可以将 $i$ 的需求进一步划分为几个子集，并确定每个子集的托管决策。

数据中心网络设计的基础模型（base model）如下所示 [见式 (4-1a)～式 (4-1k)]：

$$\min_{\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z},\ \tilde{\boldsymbol{x}},\ \tilde{\boldsymbol{z}}} \quad \underbrace{\sum_{j\in\mathcal{J}} f_j x_j + \alpha \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} c_j w_k z_{jk} +}_{\text{固定成本及耗能成本}}$$

$$\underbrace{\sum_{i\in\mathcal{I},\ j\in\mathcal{J}} d_i t_{ij} y_{ij} + \sum_{j\in\mathcal{J}} L_j(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot}) +}_{\text{传输延迟与终端延迟成本}}$$

$$\underbrace{\sum_{i\in\mathcal{I}} \tilde{f}_i \tilde{x}_i + \alpha \sum_{i\in\mathcal{I},\ k\in\mathcal{K}} \tilde{c}_i w_k \tilde{z}_{ik} + \sum_{i\in\mathcal{I}} L_i(\tilde{x}_i,\ \tilde{\boldsymbol{z}}_{i\cdot})}_{\text{托管中心成本}} \tag{4-1a}$$

$$\text{s.t.} \quad y_{ij} \leqslant x_j, \quad \forall i \in \mathcal{I},\ \ j \in \mathcal{J} \tag{4-1b}$$

$$\sum_j y_{ij} + \tilde{x}_i \geqslant 1, \quad \forall i \in \mathcal{I} \tag{4-1c}$$

$$\sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij} \leqslant z_{jk}, \quad \forall j \in \mathcal{J},\ \ \forall k \in \mathcal{K} \tag{4-1d}$$

$$\sum_{k\in\mathcal{K}} w_k z_{jk} \leqslant p_j, \quad \forall j \in \mathcal{J} \tag{4-1e}$$

$$z_{jk} \leqslant \bar{r}_{kl} z_{jl}, \quad \forall j \in \mathcal{J},\ \ \forall k,\ l \in \mathcal{K} \tag{4-1f}$$

$$d_i u_{ik} \tilde{x}_i \leqslant \tilde{z}_{ik}, \quad \forall i \in \mathcal{I},\ \ \forall k \in \mathcal{K} \tag{4-1g}$$

$$\sum_k w_k \tilde{z}_{ik} \leqslant \tilde{p}_i, \quad \forall i \in \mathcal{I} \tag{4-1h}$$

$$\tilde{z}_{ik} \leqslant \bar{r}_{kl} \tilde{z}_{il}, \quad \forall i \in \mathcal{I},\ \ \forall k,\ l \in \mathcal{K} \tag{4-1i}$$

$$x_j,\ \tilde{x}_i,\ y_{ij} \in \{0,\ 1\}, \quad \forall i \in \mathcal{I},\ \ \forall j \in \mathcal{J} \tag{4-1j}$$

$$z_{jk} \geqslant 0,\ \tilde{z}_{ik} \geqslant 0 \quad \forall j \in \mathcal{J},\ \ \forall i \in \mathcal{I},\ \ \forall k \in \mathcal{K} \tag{4-1k}$$

其中，$w_k$ 表示单位资源 $k$ 的峰值能源消耗，$\alpha$ 表示峰值能源消耗占比。模型的目标函数 (4-1a)包括了建设数据中心的固定成本、平均能量消耗成本、网络传输延迟成本和终端延迟成本。在基础模型中，假设了能源消耗成本与消耗的能源成线性比例关系，后文将放松这一假设，考虑非线性的成本结构。约束 (4-1b)和约束 (4-1c)保证了每个需求点恰好被一个选中的数据中心所服务，约束 (4-1d)和约束(4-1g)保证了数据中心的资源足以满足分配到该中心的总需求，约束 (4-1e)和约束(4-1h)保证了数据中心资源所消耗的能源在上限以内，约束(4-1f)和约束(4-1i)限制了同一数据中心中不同资源之间的最大及最小比值。

上文构造了一个从零开始布局整个数据中心网络的静态模型。在实践中，数据中心网络可能需要随时间动态扩展或重新设计。动态模型通常很难解决，而且这种长期决策所面临的不确定性进一步限制了它们的适用范围。可以对以上模型进行拓展以解决现有网络的网络扩展或重新设计问题。如果需要考虑数据中心的一些其他特性，例如关闭现有数据中心、移动或回收部署的服务器，以及迁移计算资源等，可以通过滚动时域法应用本模型，以提供网络扩展和重新配置计划。

从建模的角度来看，与托管相关的所有决策与数据中心的决策都没有结构上的差异。因此，为清楚起见，在随后的分析中省略了托管中心的决策，但将在数值研究中重新探讨托管中心。下一个小节将进一步刻画求解目标函数中的终端延迟成本项 $L_j(\boldsymbol{y}_{\cdot j}, \boldsymbol{z}_{j\cdot})$。

### 4.3.2　终端延迟成本

排队模型已广泛应用于计算机系统性能的建模，例如，Harchol-Balter（2013）总结了常见的计算机系统中可应用的排队模型。本小节用数据包在数据中心所经历的平均逗留时间来衡量终端延迟，并分析排队模型以估算终端延迟成本。

为了便于分析，做出以下假设。首先，假设数据中心在处理器共享（processor sharing）准则下运行，这样任何传入的作业都会立即被调度以进行处理，并且总资源将被平均分配到每个传入的作业中。处理器共享已成为云计算文献中使用最广泛的范例之一（Altman et al.，2011）。实际上，类似谷歌这样的业界巨头使用的就是类似于处理器共享的管理策略

（Silberschatz et al.，1998；Verma et al.，2015）。目前也存在一些高级的调度和优先级策略用以降低终端延迟，然而作为一个战略层面规划模型，不必详细刻画每个作业的调度，而是使用处理器共享策略来近似平均的终端等待时间。

处理一个作业需要几种不同的资源，每个作业一次只占用一种类型的资源。例如，几乎所有的计算机程序都需要交替使用计算资源和存储输入输出资源（Silberschatz et al.，1998）。用服务阶段来区分同一个作业消耗不同资源的时段，并进一步假设服务阶段形成了一个串联的队列。因此，假设作业在每个阶段的服务时间是互相独立的，数据中心处理一个作业所需要的时间包括了作业在不同服务阶段的等待时间之和。附录 C.1.1 考虑了更一般的情况，将串联的队列拓展到了排队网络（open queueing network）中。排队网络包括计算和存储服务两个服务阶段，在最终离开网络之前，作业以交替的方式随机路由到这两个服务阶段。为了清楚起见，在本书的其他部分中，将仍然遵循串联队列假设，但所有理论结果和解决方法均仍适用于排队网络。

本书通过使用具有不同平均服务时间的多个客户类别来建模需求异质性。假设目前数据中心 $j$ 仅服务于需求点 $i$，如果数据中心 $j$ 处的资源 $k$ 保有量为 $z_{jk}$，则数据中心 $j$ 满足需求点 $i$ 阶段 $k$ 的需求的平均服务率为 $z_{jk}/u_{ik}$。进一步假设阶段服务时间遵循 Coxian 分布。图 4-2 展示了如果每个阶段的服务时间使用 Coxian-2 分布来近似，那么数据中心中具有两个阶段的串联队列的停留时间。其中 Coxian-2 分布的参数 $(\mu^i_{c1},\ \mu^i_{c2},\ p^i_c)$ 和 $(\mu^i_{s1},\ \mu^i_{s2},\ p^i_s)$ 是通过与服务时间的前三阶矩拟合得到的。这些参数是资源供应决策 $\boldsymbol{z}$ 的函数，这对附录 C.1.1 中排队模型关键参数的推导有重要作用。实际上 Coxian 分布的假设并没有对延迟时间的估计进行过多限制，因为对于任何具有 Laplace-Stieltjes 变换的分布

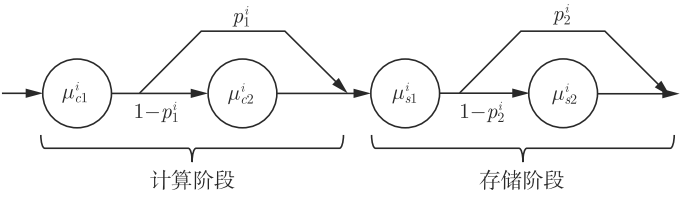



**图 4-2　两个服务阶段的串联队列**

注：服务等待时间满足 Coxian-2 分布，有两个服务阶段的串联队列对平均等待时间的近似。

来说，Coxian 分布都可以通过矩匹配来近似任意分布（Cox，1955）。

在以上的假设条件下，每个数据中心包括有 $|\mathcal{K}|$ 个阶段的串联队列，每个阶段包含着多类别，但服务器的处理器共享排队模型，其服务时间服从 Coxian 分布。在命题 4-1 中，用特定阶段平均等待时间来估计终端延迟成本。值得注意的是，通过使用更广泛的 phase-type 分布，可以将后续分析推广到具有不同资源替代占用的情况。

**命题 4-1：** 平均等待时间。

作业在数据中心 $j$ 的平均等待时间 $W_j(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot})$，如下所示：

$$W_j(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot}) = \sum_{k\in\mathcal{K}} \frac{\mathbb{E}[S_{jk}]}{1-\lambda_j \mathbb{E}[S_{jk}]}$$

其中 $\mathbb{E}[S_{jk}] = \dfrac{\sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij}}{z_{jk}\sum_{i\in\mathcal{I}} d_i y_{ij}}$ 表示在分配到数据中心 $j$ 的所有需求中，作业在数据中心 $j$ 的阶段 $k$ 的期望服务时间，$\lambda_j = \sum_i d_i y_{ij}$ 表示数据中心 $j$ 的需求到达率。

此外，从需求点 $i$ 分配到数据中心 $j$ 的作业等待时间 $W_j^i(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot})$ 满足：

$$W_j^i(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot}) = \sum_{k\in\mathcal{K}} \frac{u_{ik} y_{ij}}{z_{jk}(1-\lambda_j \mathbb{E}[S_{jk}])}$$

命题 4-1 表明数据中心服务阶段的作业平均等待时间与该服务阶段的已调配资源量成反比，与服务阶段空闲时间百分比成反比。命题 4-1 的证明受 Baskett et al.（1975）开创性的工作启发。同时，附录 C.1.1 中简要讨论了对于一般 phase-type 分布下的等待时间对命题 4-1 的拓展。

命题 4-1 的第二部分为计算异质顾客的终端等待时间提供了基础。具体来说，本模型考虑了不同顾客对终端等待时间的异质性，并用单位终端延迟成本 $\tau_i$ 表示。给定需求分配决策 $\boldsymbol{y}_{\cdot j}$ 和资源供给决策 $\boldsymbol{z}_{j\cdot}$，数据中心 $j$ 所带来的总终端延迟成本的期望 $L_j(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot})$ 可表示如下 [见式 (4-2)]：

$$L_j(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot}) = \sum_i \tau_i d_i W_j^i(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot}) = \sum_{k\in\mathcal{K}} \frac{\sum_{i\in\mathcal{I}} \tau_i d_i u_{ik} y_{ij}}{z_{jk} - \sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij}} \tag{4-2}$$

将式(4-2) 代入基础模型中，可以得到下述考虑异质顾客的数据中心网络设计模型：

$$(\mathbf{P}) \quad \min_{\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z}} \sum_{j\in\mathcal{J}} f_j x_j + \alpha \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} c_j w_k z_{jk} + \sum_{i\in\mathcal{I},\ j\in\mathcal{J}} d_i t_{ij} y_{ij} + \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} \frac{\sum_{i\in\mathcal{I}} \tau_i d_i u_{ik} y_{ij}}{z_{jk} - \sum_i d_i u_{ik} y_{ij}}$$

s.t. 约束式(4-1b) ~ 式(4-1f)， 式(4-1j) ~ 式(4-1k)

问题（P）是一个混合整数非线性（MINLP）规划模型，其中目标函数中包含有线性分式。该问题可以很容易从无容量限制的设施选址问题归约为 NP-hard 问题。附录 C.2 还讨论了该模型的可近似性。在介绍该模型的解法前，在下面的几个小节对本模型进行几个重要的拓展。

### 4.3.3 需求点的相互依赖性

首先考虑需求点之间的相互依赖性。在云计算中，有时需要将作业从其指定的数据中心重新路由到不同的数据中心，以提供进一步的服务。这样的路由可能是由于需求点之间需要同步，冗余或额外的数据访问导致的（Greenberg et al.，2008）。

在不失一般性的前提下，假设一部分作业在其指定数据中心的服务完成后路由到另一个数据中心以进行其他服务，然后离开系统。假设从需求点 $i$ 路由到需求点 $i'$ 的专用数据中心以进行进一步服务的概率为 $P_{ii'}$。因此，数据中心 $j$ 的来源与需求点 $i$ 需求到达率由式 (4-3) 给出：

$$\lambda_{ij} = d_i y_{ij} + \sum_{i'\neq i} d_i P_{ii'} y_{i'j} \triangleq d_i \sum_{i'} P_{ii'} y_{i'j} \tag{4-3}$$

其中，对于所有 $i \in \mathcal{I}$，有 $P_{ii} \triangleq 1$。注意到 $P_{ii'}$ 与排队网络中的路由概率并不相同，$\sum_{i'\neq i} P_{ii'}$ 可以不等于 1。例如，一个需求点 $i$ 的作业可能需要其他多个数据中心对其服务才能满足，因此 $\sum_{i'\neq i} P_{ii'} > 1$。

需要附加服务的每个作业会产生两个额外费用，分别是与附加服务相关的额外终端主机延迟成本，以及在不同数据中心处理附加服务时数据中心之间的额外传输延迟成本。考虑到异质延迟成本的数据中心 $j$ 的总终端延迟成本见式 (4-4)：

$$L_j(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot}) = \sum_k \frac{\sum_i \left(\sum_{i'} \tau_{i'} d_{i'} u_{i'k} P_{i'i}\right) y_{ij}}{z_{jk} - \sum_i \left(\sum_{i'} d_{i'} u_{i'k} P_{i'i}\right) y_{ij}} \tag{4-4}$$

给定来自需求点 $i$ 从数据中心 $j$ 路由到数据中心 $j'$ 的需求强度为 $d_i y_{ij} \sum_{i'} P_{ii'} y_{i'j'}$，两个数据中心之间的传输延迟成本可以表示为式 (4-5)：

$$\sum_{i,\ i'\in\mathcal{I}} \sum_{j,\ j'\in\mathcal{J}} \psi_i t_{jj'} d_i P_{ii'} y_{ij} y_{i'j'} \tag{4-5}$$

其中，$t_{jj'}$ 表示数据中心 $j$ 到 $j'$ 之间的网络传输延迟成本，$\psi_i$ 表示从需求点 $i$ 路由到其他数据中心的延迟成本校正系数。模型可以很容易拓展到作业大小在路由后发生变化的情况。令 $\Gamma_{ii'}$ 表示从需求点 $i$ 到 $i'$ 的作业变化校正系数，其中 $\Gamma_{ii}=1$。数据包大小的变化只会影响额外的终端延迟成本，具体如下所示：

$$L_j(\boldsymbol{y}_{\cdot j},\ \boldsymbol{z}_{j\cdot}) = \sum_k \frac{\sum_i \left(\sum_{i'} \tau_{i'} d_{i'} u_{i'k} P_{i'i} \Gamma_{i'i}\right) y_{ij}}{z_{jk} - \sum_i \left(\sum_{i'} d_{i'} u_{i'k} P_{i'i} \Gamma_{i'i}\right) y_{ij}}$$

为表示方便，暂时忽略数据包的大小变化。将式(4-5) 加入目标函数，并将终端延迟成本用式(4-4)表示，可以得到下述考虑数据中心之间依赖关系的数据中心网络设计模型 [见式 (4-6)、式 (4-7)]：

$$\textbf{(P-ID)}\quad \min_{\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z}} \quad \sum_{j\in\mathcal{J}} f_j x_j + \alpha \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} c_j w_k z_{jk} + \sum_{i\in\mathcal{I},\ j\in\mathcal{J}} d_i t_{ij} y_{ij} + \tag{4-6}$$

$$\sum_{i,\ i'\in\mathcal{I}} \sum_{j,\ j'\in\mathcal{J}} \psi_i t_{jj'} d_i P_{ii'} y_{ij} y_{i'j'} + \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} \frac{\sum_i \left(\sum_{i'} \tau_{i'} d_{i'} u_{i'k} P_{i'i}\right) y_{ij}}{z_{jk} - \sum_i \left(\sum_{i'} d_{i'} u_{i'k} P_{i'i}\right) y_{ij}}$$

$$\text{s.t.}\quad \sum_i \left(\sum_{i'} d_{i'} u_{i'k} P_{i'i}\right) y_{ij} \leqslant z_{jk}, \quad \forall j\in\mathcal{J},\ k\in\mathcal{K} \tag{4-7}$$

约束式(4-1b)～式(4-1c)，式(4-1e)～式(4-1f)，式(4-1j)～式(4-1k)

问题（P-ID）是单分配的轴幅网络选址问题（single-allocation hub location problem）的一个拓展，在计算上比普通 MINLP 的问题更加困难。

### 4.3.4　凸电能消耗

研究表明，数据中心的高利用率可能会导致用电效率（PUE）降低。因此电能消耗通常是使用率的递增的凸函数（Chen et al.，2013）。为了

表示利用率提高所带来的用电效率下降的情况，假设数据中心 $j$ 的每单位资源 $k$ 的平均功耗为 $\omega_{jk}$，是一个对于使用率（utilization）$\rho_{jk}$ 的多项式函数类型的一般凸函数，其中 $\rho_{jk}=\sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij}/z_{jk}$。换言之，即式 (4-8)：

$$\omega_{jk}=\sum_{l} a_{jk}^{l}\rho_{jk}^{\sigma^l}+b_{jk},\qquad \rho_{jk}=\sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij}/z_{jk} \tag{4-8}$$

其中 $a_{jk}^{l}$，$b_{jk}$，$\sigma^l$ 这些参数可以从历史电能消耗数据中估计出来，$\sigma^l$ 可以是任意大于等于 1 的有理数，甚至可以对于不同的数据中心、不同的资源，取值不同（如果 $\sigma^l$ 取正整数，那么 $\omega_{jk}$ 则退化成一个多项式函数）。此外，假设 $a_{jk}^{l}\geqslant 0$，$b_{jk}\geqslant 0$，数据中心的电量消耗成本是非负的，数据中心 $j$ 使用资源 $k$ 所消耗的平均电量成本为 $c_j z_{jk}\omega_{jk}$。为了在目标函数中将这一项线性化，用 $\alpha\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} c_j w'_{jk}$ 来做替换，并且加入下列约束[见式 (4-9)]：

$$w'_{jk}\geqslant \omega_{jk} z_{jk} \tag{4-9}$$

代入约束式(4-8)和式(4-9)，考虑凸电能消耗的模型如下所示：

$$\textbf{（P-CP）}\quad \min_{\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z}} \sum_{j\in\mathcal{J}} f_j x_j+\alpha\sum_{j\in\mathcal{J},\ k\in\mathcal{K}} c_j w'_{jk}+\sum_{i\in\mathcal{I},\ j\in\mathcal{J}} d_i t_{ij} y_{ij}+\sum_{j\in\mathcal{J},\ k\in\mathcal{K}} \frac{\sum_{i\in\mathcal{I}}\tau_i d_i u_{ik} y_{ij}}{z_{jk}-\sum_i d_i u_{ik} y_{ij}}$$

s.t. 约束式(4-1b) ~ 式(4-1f)，式(4-1j) ~ 式(4-1k)，式(4-8)，式(4-9)

值得注意的是，约束式(4-8)包含指数多项式函数，并且约束式(4-9)需要进一步地线性化，这进一步提高了问题（P-CP）的复杂度。最后，尽管文献中广泛使用二次函数来拟合凸电能消耗，相比之下，本部分提出的模型允许指数 $\sigma^l$ 取大于或等于 1 的任何有理数，能够更好地拟合电能消耗函数。

### 4.3.5 网络拥堵

大量的数据交换可能会导致数据中心网络的网络拥堵。网络拥堵相关的理论模型一般假设传输延迟是网络流量强度的单调递增凸函数，而且与线缆的承载能力相关（Spiess，1990；Luna & Mahey，2000）。

为了刻画网络拥堵，本部分采用了美国公共道路局（U.S. Bureau of Public Roads）研究出的 BPR 函数。具体来说，从需求点 $i$ 到数据中心

$j$ 的传输延迟成本为：

$$d_i y_{ij} t_{ij} \left[1 + \left(\frac{d_i}{\chi_{ij}}\right)^{\sigma_{ij}}\right]$$

其中 $\chi_{ij}$ 是一个新的决策变量，用来表示从需求点 $i$ 到数据中心 $j$ 的线缆承载能力，参数 $\boldsymbol{\sigma}$ 可以通过历史数据估计得到。假设 $\sigma_{ij} > 1$ 来保证传输延迟是网络流量强度的单调递增凸函数。为了使得目标函数变为线性函数，用 $s_{ij}$ 来表示网络传输延迟，并加入以下约束 [见式 (4-10)]：

$$s_{ij} \geqslant y_{ij} + \left(\frac{d_i}{\chi_{ij}}\right)^{\sigma_{ij}} y_{ij} \tag{4-10}$$

如果考虑需求之间的互相影响，数据中心之间的拥堵程度也可以用类似的方法表示。具体来说，从数据中心 $j$ 到数据中心 $j'$ 的网络传输延迟成本为：

$$t_{jj'} d_{jj'} \left[1 + \left(\frac{d_{jj'}}{\chi_{jj}}\right)^{\sigma_{jj'}}\right]$$

其中 $d_{jj'} = \sum_{i,\ i'} \psi_i d_i y_{ij} P_{ii'} y_{i'j'}$，表示从数据中心 $j$ 到数据中心 $j'$ 的数据流量强度。使用 $s_{jj'}$ 来表示数据中心之间的网络传输延迟，并加入以下约束 [见式 (4-11)]：

$$s_{jj'} \geqslant d_{jj'} + \frac{d_{jj'}^{\sigma_{jj'}+1}}{\chi_{jj'}^{\sigma_{jj'}}} \tag{4-11}$$

如果同时加入约束式 (4-10) 和式 (4-11)，可以得到以下包含所有拓展的模型：

$$\begin{aligned}
\textbf{(P-CC)}\quad \min_{\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z},\ \boldsymbol{\chi}} \ & \sum_{j\in\mathcal{J}} f_j x_j + \alpha \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} c_j w'_{jk} + \sum_{i\in\mathcal{I},\ j\in\mathcal{J}} (d_i t_{ij} s_{ij} + \kappa_{ij}\chi_{ij}) + \\
& \sum_{jj'\in\mathcal{I},\ j\in\mathcal{J}} (d_{jj'} s_{jj'} + \kappa_{jj'}\chi_{jj'}) + \\
& \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} \frac{\sum_i \left(\sum_{i'} \tau_{i'} d_{i'} u_{i'k} P_{i'i}\right) y_{ij}}{z_{jk} - \sum_i \left(\sum_{i'} d_{i'} u_{i'k} P_{i'i}\right) y_{ij}} \\
\text{s.t.}\quad & \text{约束式(4-1b)} \sim \text{式(4-1c)，式(4-1e)} \sim \text{式(4-1f)，} \\
& \text{式(4-1j)} \sim \text{式(4-1k)，式(4-7)} \sim \text{式(4-11)}
\end{aligned}$$

其中 $\kappa_{ij}$ 和 $\kappa_{jj'}$ 表示线缆承载能力的单位建造成本。本书注意到约束式(4-10)和式(4-11) 是混合整数非线性约束，这为问题的求解带来了进一步的挑战。

## 4.4 模型解决方法

在本节中，首先构造原模型的一些等价变形，并分析其结构特性。然后，开发了拉格朗日松弛法对变形后的模型进行求解。最后，在拉格朗日松弛法的基础上加入了 extremal extended polymatroid 割平面，进一步加快求解速度。

### 4.4.1 等价变形

前述所有模型及其拓展都可以等价变形为混合整数二阶锥优化（MISOCP）。首先证明问题（P）的目标函数中的线性分式项是可以用二阶锥表示的。这一结果可以正式表述为如下命题：

**命题 4-2：** 终端延迟的等价变形。

问题 （P）（考虑异质需求的基础数据中心网络设计问题）与下述问题 $(\bar{\mathrm{P}})$ 等价 [见式 (4-12)]：

$$
\begin{aligned}
\min_{\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z},\ \boldsymbol{v}} \ & \sum_{j\in\mathcal{J}} f_j x_j + \alpha \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} c_j w_k z_{jk} + \sum_{i\in\mathcal{I},\ j\in\mathcal{J}} d_i t_{ij} y_{ij} + \sum_{j\in\mathcal{J},\ k\in\mathcal{K}} v_{jk} \\
\text{s.t.} \ & \left\| \begin{pmatrix} 2\Lambda_k \boldsymbol{y}_{\cdot j} \\ v_{jk} - z_{jk} + \sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij} \end{pmatrix} \right\|_2 \leqslant v_{jk} + z_{jk} - \sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij}, \\
& \quad \forall j \in \mathcal{J},\ k \in \mathcal{K} \qquad (4\text{-}12) \\
& v_{jk} \geqslant 0, \quad \forall j \in \mathcal{J},\ k \in \mathcal{K} \\
& \text{约束式(4-1b)} \sim \text{式(4-1k)}
\end{aligned}
$$

其中 $\Lambda_k \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ 是一个对角阵，对角元素 $(\Lambda_k)_{ii} = \sqrt{\tau_i d_i u_{ik}}$。问题 $(\bar{\mathrm{P}})$ 是一个 MISOCP 问题。

变形后的问题 $(\bar{\mathrm{P}})$ 的目标函数是线性的，其约束条件都是二阶（凸）锥约束，因此可以使用求解器对变形的问题直接求解。然而当问题的规模比较大的时候，这些求解器的效率就会变得比较低。

接下来，在研究不同需求点的依赖性的模型（P-ID）中，网络传输延迟成本式(4-5) 可以被线性化。过去的文献中有很多线性化的手段，包

括 Balas & Mazzola（1984）提出的标准方法、紧缩的标准方法和最近在 0-1 二次规划中的一种紧变形（Chaovalitwongse et al.，2004；Sherali & Smith，2007），这种方法只需要线性数量的辅助变量和约束来线性化。在紧变形方法中，引入了辅助变量 $w_{ij}$ 和 $v_{ij}$，同时把目标函数中的中转网络传输成本替换为 $\sum_{i\in\mathcal{I},\ j\in\mathcal{J}} w_{ij}$ 并加入以下约束 [见式 (4-13a)、式 (4-13b)、式 (4-13c)]：

$$\sum_{i'}\sum_{j'} d_i\psi_i t_{jj'}P_{ii'}y_{i'j'} - v_{ij} = w_{ij}, \quad \forall i\in\mathcal{I},\ \forall j\in\mathcal{J} \tag{4-13a}$$

$$v_{ij} \leqslant \sum_{i'}\sum_{j'} t_{jj'}d_i\psi_i P_{ii'}(1-y_{ij}), \quad \forall i\in\mathcal{I},\ \forall j\in\mathcal{J} \tag{4-13b}$$

$$w_{ij} \geqslant 0, \quad \forall i\in\mathcal{I},\ \forall j\in\mathcal{J} \tag{4-13c}$$

在考虑凸电能消耗的问题（P-CP）中，非凸的约束式(4-8) 和式(4-9)也可以转化成二阶锥约束。

**命题 4-3：** 凸电能消耗的等价变形。

在问题 （P-CP）中，约束式 (4-8)和式(4-9)与下列约束 [见式 (4-14)、式 (4-15)] 等价：

$$w'_{jk} \geqslant \sum_{l} a^l_{jk}\tilde{w}^l_{jk} + b_{jk}z_{jk}, \quad \forall j\in\mathcal{J},\ k\in\mathcal{K} \tag{4-14}$$

$$\tilde{w}^l_{jk} z^{\sigma^l-1}_{jk} \geqslant \left(\sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij}\right)^{\sigma^l}, \quad \forall j\in\mathcal{J},\ k\in\mathcal{K},\ \forall l \tag{4-15}$$

其中式 (4-14) 是一个线性约束，而式(4-15) 是线性约束（当 $\sigma^l = 1$）或者可以表示为二阶锥约束（当 $\sigma^l$ 是一个大于 1 的有理数）。

命题 4-3 说明了非凸的约束式 (4-8) 和式 (4-9) 是可以通过二阶锥约束表示出来的。本书的其余部分，为了简洁起见，假设电能消耗是使用率的二次函数。

最后，在考虑网络拥堵的问题 （P-CC）中，非凸约束式(4-10) 和式(4-11)也可以表示为二阶锥约束。

**命题 4-4：** 网络拥堵的等价变形。

在问题 （P-CC）中，约束式(4-10) 等价于式 (4-16)：

$$s_{ij} \geqslant y_{ij} + \pi_{ij}$$

$$d_i^{\sigma_{ij}} y_{ij}^{\sigma_{ij}+1} \leqslant \chi_{ij}^{\sigma_{ij}} \pi_{ij} \tag{4-16}$$

而约束式 (4-11) 等价于式 (4-17)：

$$\begin{aligned} s_{jj'} &\geqslant d_{jj'} + \pi_{jj'} \\ d_{jj'}^{\sigma_{jj'}+1} &\leqslant \chi_{jj'}^{\sigma_{jj'}} \pi_{jj'} \end{aligned} \tag{4-17}$$

当 $\sigma_{ij}$ ($\sigma_{jj'}$) 是正有理数的时候，两约束式(4-16)和式(4-17) 都可以转化为二阶锥约束。

综上所述，问题（P）及其所有拓展都可以变形成为等价的 MISOCP 问题，并可以直接使用 CPLEX、Gurobi 等商业软件求解。但是直接求解现实中的算例仍然需要使用计算能力很强的计算机，即便如此有的时候也无法在有限时间内找到可行解。

在给出具体的解决方案之前，本书注意到终端延迟成本 $L_j(\boldsymbol{y}_{\cdot j}, \boldsymbol{z}_{j\cdot})$ 在之前所有的模型中都起到了重要影响。注意到一个作用在集合上的函数 $g(\cdot)$ 被称作超模（supermodular），如果对于任意两个集合 $S$， $T \subset \mathcal{V}$，有 $g(S \cup T) + g(S \cap T) \geqslant g(S) + g(T)$。这一超模概念也可以拓展到网格域中。下面的命题 4-5 说明了 $L_j(\boldsymbol{y}_{\cdot j}, \boldsymbol{z}_{j\cdot})$ 具有超模性。

**命题 4-5：** 终端延迟的超模性。

对于每个数据中心 $j \in \mathcal{J}$，定义函数 $\tilde{L}_j : \{0, 1\}^{|\mathcal{I}|} \times \mathbb{R}_-^{|\mathcal{K}|} \to \mathbb{R}_+ \cup \{+\infty\}$ 如下：

$$\tilde{L}_j(\boldsymbol{y}_{\cdot j}, \boldsymbol{z}_{j\cdot}^-) = \begin{cases} L_j(\boldsymbol{y}_{\cdot j}, \boldsymbol{z}_{j\cdot}), & \text{if } \tilde{z}_{jk} + \sum_i d_i u_{ik} y_{ij} < 0, \quad \forall k \in \mathcal{K} \\ +\infty, & \text{其他} \end{cases}$$

其中 $\boldsymbol{z}^- = -\boldsymbol{z}$。函数 $\tilde{L}_j$ 关于 $(\boldsymbol{y}_{\cdot j}, \boldsymbol{z}_{j\cdot}^-)$ 在 $\{0, 1\}^{|\mathcal{I}|} \times \mathbb{R}_-^{|\mathcal{K}|}$ 上是单调函数且具有超模性。

命题 4-5 揭示了需求分配和资源供应决策的边际影响。该命题有以下两个方面的意义：（1）由于数据中心的总电能的物理上限和有限的预算约束，一些数据中心无法继续拓展。如果把需求点分配到快达到容量极限的数据中心，从终端延迟角度来看会带来更高的成本，因此应该将需求分配给负载最轻的数据中心。（2）可能由于传输延迟成本所致，需求分配决策难以调控的时候，如果有额外的预算来投资资源调配，应该投资于负

载较重的数据中心。利用这些特性可以为提出的拉格朗日松弛方法设计伴随的启发式方法，该方法将在下一部分中介绍。

### 4.4.2　拉格朗日松弛算法

拉格朗日松弛算法能够有效解决大规模有容量约束的选址（CFL）问题。拉格朗日松弛算法一般有两种应用的方法——放松容量约束式(4-1e)或放松需求分配约束式(4-1c)（Beasley，1993；Sridharan，1995）。如果放松需求分配约束，CFL 的子问题可以转化为背包问题，但是在本部分的模型下有比较复杂的终端延迟成本，使用放松需求分配约束的方法并不能对模型求解速度有太大的帮助。因此，本节放松了容量约束(4-1e)来求解 MISOCP 问题 ($\bar{\mathrm{P}}$)。值得指出的是，该方法也适用于其他模型拓展。令 $\boldsymbol{\lambda} \in \mathbb{R}_{+}^{|\mathcal{J}|}$ 为约束式(4-1e)对应的拉格朗日乘子。放松了该约束后，拉格朗日子问题也是一个 MISOCP 问题，并且可以表示成如下式子：

$$\textbf{（P-L）}\quad Z_D(\lambda) = \min_{\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z},\ \boldsymbol{v}} \sum_j f_j x_j + \alpha \sum_{j,\ k} c_j w_k z_{jk} + \sum_{i,\ j} d_i t_{ij} y_{ij} + \sum_{j,\ k} v_{jk} + \sum_j \lambda_j \left( \sum_k w_k z_{jk} - p_j \right)$$

$$\text{s.t. (4-1b)}\sim\text{(4-1d)，(4-1f)，(4-1j)，(4-1k)(4-12)}$$

直观上来说，拉格朗日子问题（P-L）相对原问题会更容易解决，因为不需要考虑资源供应的约束上限。给定拉格朗日乘子 $\lambda^{(m)}$，其中上角标 $(m)$ 表示迭代次数，拉格朗日子问题的目标函数值 $Z_D(\lambda^{(m)})$ 提供了问题（P）的一系列下界。如果拉格朗日子问题的解对于原问题（P）也是可行的，那么这个解则对应原问题的一个上界 $\bar{Z}$。当 $\bar{Z}$ 和 $Z_D(\lambda^{(m)})$ 足够接近的时候，则终止算法，得出最优解。本部分使用文献中常见的规则来更新拉格朗日乘子 $\lambda$[具体可参考 Fisher（1981）]。为了表示方便，在接下来的讨论中，将省略掉迭代次数的角标 $(m)$。

由于拉格朗日子问题（P-L）的解不一定总对原问题（P）是可行的，也无法保证上界 $\bar{Z}$ 的持续更新。为了加速收敛，接下来提出了一个启发式算法来把问题（P-L）的解转化成原问题的可行解。算法的核心想法就是找到超过容量限制的数据中心，选出一些分配到该数据中心的需求点，并把这些需求点交换到其他未到容量限制的数据中心中去。

展开来讲，对于每个原问题不可行的解，按照容量约束将数据中心分成三个集合。集合 $A$ 包括超过容量限制的数据中心，集合 $B$ 包括被选中的未超过容量限制的数据中心，集合 $C$ 包括未被选中的数据中心。该启发式算法将集合 $A$ 中的数据中心所分配的需求点逐个重新分配到集合 $B$ 中，如果集合 $B$ 中没有足够的容量来服务需求，那么就从集合 $C$ 中选出一个新的数据中心。启发式算法会持续交换需求点分配，直到集合 $A$ 变为空集。

启发式算法如算法 (4-1) 所示。假设在拉格朗日子问题（P-L）的最优解中，需求点 $s$ 被分配到一个超过容量限制的数据中心 $t$ 中，交换 $s$ 的服务数据中心。算法 4-1 中的第 7 行和第 16 行是从集合 $B$ 或 $C$ 中选出候选数据中心的关键步骤。交换步骤的核心在于估计每次交换带来的成本提升。用 $\delta(s，t，t')$ 来表示将需求点从数据中心 $t$ 交换到数据中心 $t'$ 的成本增加的上限。由于 $\delta(s，t，t')$ 保证了对偶间隙的上界，命题 4-6 可以用以检查现有的可行解是否与最优解足够接近。

**命题 4-6：** 现有可行解的界限。

令 $z_{tk}$，$z_{t'k}$（$\forall k \in \mathcal{K}$）分别为从数据中心 $t$ 交换到数据中心 $t'$ 之前的资源供给水平。假设数据中心 $t'$ 在被交换后没有超出容量限制，那么 $\delta(s，t，t')$ 满足：

（a）如果 $t' \in B$，那么 $\delta(s，t，t')$ 的上界为：

$$\delta(s，t，t') = d_s \left( \tau_s \sum_k u_{sk} \left( \frac{1}{z_{t'k} - \sum_i d_i u_{ik} y_{it'}} - \frac{1}{z_{tk} - \sum_i d_i u_{ik} y_{it}} \right) + \alpha(c_{t'} - c_t) \sum_k u_{sk} w_k + t_{st'} - t_{st} \right)$$

（b）如果 $t' \in C$ 而且 $t'$ 的容量限制适中，满足 $p_{t'} \geqslant \sum_k w_k (z_{tk} - \sum_i d_i u_{ik} y_{it})$，那么 $\delta(s，t，t')$ 的上界满足：

$$\delta(s，t，t') = f_{t'} + \alpha c_{t'} \sum_k w_k \left( z_{tk} - \sum_i d_i u_{ik} y_{it} \right) - d_s \left( \alpha c_t \sum_k u_{sk} w_k - t_{st'} + t_{st} \right)$$

鉴于命题 4-5 中的超模性，本部分的启发式算法在选择新的数据中心的时候，首先会按照算法 (4-1) 第 8 行和第 17 行中的策略来估计数据中心是否有足够的容量并且资源比例约束是否足够松。每次交换步骤之后，算法 4-1 的第 24 行会重新优化两个数据中心的总资源供给决策，并计算出此时的总终端延迟成本。（Re-optimize）步骤计算了给定需求分配决策后的，满足容量约束和资源比例约束下的最优资源供给水平。该步骤求解了以下问题：

$$
\begin{aligned}
(\textbf{Re-optimize}) \qquad \min_{\boldsymbol{z}_{j\cdot},\ \boldsymbol{v}_j} \quad & \sum_k \alpha c_j w_k z_{jk} + v_{jk} \\
\text{s.t.} \quad & \left\| \begin{pmatrix} 2\Lambda_k \boldsymbol{y}_{\cdot j} \\ v_{jk} - z_{jk} + \sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij} \end{pmatrix} \right\|_2 \\
& \leqslant v_{jk} + z_{jk} - \sum_{i\in\mathcal{I}} d_i u_{ik} y_{ij}, \quad \forall k \in \mathcal{K} \\
& z_{jk} \leqslant \bar{r}_{kl} z_{jl}, \qquad \forall k,\ l \in \mathcal{K} \\
& \sum_{k\in\mathcal{K}} w_k z_{jk} \leqslant p_j
\end{aligned}
$$

该子问题是一个只有连续决策变量的二阶最优化问题，使用商业软件就可以高效解决。

**算法 4-1:** 问题（P）的可行解的启发式算法

1: **初始化** $\boldsymbol{A} = \{j : x_j = 1,\ \sum_k w_k z_{jk} > p_j\}$，$\boldsymbol{B} = \{j : x_j = 1,\ \sum_k w_k z_{jk} \leqslant p_j\}$，$\boldsymbol{C} = \{j : x_j = 0\}$
2: **while** $\boldsymbol{A} \neq \varnothing$ **do**
3: $t = \arg\max_{j\in\boldsymbol{J}} \{\sum_k (w_k z_{jk}) - p_j\}$
4: $s = \arg\min_{i\in\{i|y_{it}=1\}} \sum_k d_i u_{ik} w_k$
5: $y_{st} \leftarrow 0$
6: **while** $\boldsymbol{B} \neq \varnothing$ **do**
7: $t' \leftarrow \arg\min_{j\in\boldsymbol{B}} \left\{ \tau_s \sum_k \frac{u_{sk}}{z_{jk} - \sum_i d_i u_{ik} y_{ij}} + \sum_k \alpha c_j u_{sk} w_k + t_{sj} \right\}$
8: **if** $\sum_k (w_k z_{t'k} + d_s u_{sk} w_k) \leqslant p_{t'}$，and $(d_s u_{sk} + z_{t'k}) \leqslant \bar{r}_{kl}(d_s u_{sl} + z_{t'l})$，$\forall k,\ l \in \boldsymbol{K}$ **then**
9: $y_{st'} \leftarrow 1$
10: **break**

11: **else**
12: 从集合 $\boldsymbol{B}$ 删除 $t'$
13: **end if**
14: **end while**
15: **while** $\boldsymbol{C} \neq \varnothing$ **do**
16: $t' \leftarrow \arg\min_{j \in \boldsymbol{C}} \{\alpha c_j \sum_k w_k (z_{tk} - \sum_i d_i u_{ik} y_{ij}) + d_s t_{sj} + f_j\}$
17: **if** $\sum_k (w_k (z_{tk} - \sum_i d_i u_{ik} y_{it})) \leqslant p_{t'}$ **then**
18: $y_{st'} \leftarrow 1$ and $x_{t'} \leftarrow 1$
19: **break**
20: **else**
21: 从集合 $\boldsymbol{C}$ 删除 $t'$
22: **end if**
23: **end while**
24: 更新集合 $\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{C}$，函数值 $(\boldsymbol{z}_{t\cdot}, \boldsymbol{v}_{t\cdot}) \leftarrow (\boldsymbol{y}_{\cdot t})$，和 $(\boldsymbol{z}_{t'\cdot}, \boldsymbol{v}_{t'}) \leftarrow \boldsymbol{Re\text{-}optimize}(\boldsymbol{y}_{\cdot t'})$
25: **end while**

此外，本部分还提出了另一个启发式算法来加速更新 $\boldsymbol{z}$ 和 $\boldsymbol{v}$，如算法 4-2 所示。该启发式算法首先对所有资源进行迭代，以计算没有容量约束和资源比率约束的最佳资源供给水平，然后通过截断（第 3 行）和缩放

**算法 4-2 ：** 更新 $\boldsymbol{z}_{j\cdot}$ 的启发式算法

1: [**Function**]$\boldsymbol{Re\text{-}optimize}(\boldsymbol{y}_{\cdot j})$
2: **for** $k = 1, \cdots, |\mathcal{K}|$ **do**
3: $z_{jk} = \max\left\{\min\left\{\sum_i d_i u_{ik} y_{ij} + \sqrt{\frac{\sum_{i \in \mathcal{I}} \tau_i d_i u_{ik} y_{ij}}{\alpha c_j w_k}},\ \bar{r}_{kl} z_{jl}\right\},\ \frac{z_{jl}}{\bar{r}_{lk}}\right\}$, for all $l \leqslant k$
4: **end for**
5: **if** $\sum_k w_k z_{jk} > p_j$ **then**
6: **for** $k = 1, \cdots, |\mathcal{K}|$ **do**
7: $z_{jk} \leftarrow \frac{p_j}{\sum_k w_k z_{jk}} z_{jk}$
8: **end for**
9: **end if**
10: $v_{jk} \leftarrow \frac{\sum_{i \in \mathcal{I}} \tau_i d_i u_{ik} y_{ij}}{z_{jk} - \sum_i d_i u_{ik} y_{ij}}$

（第 7 行），以确保它们满足约束条件。通过结合算法 4-1 和算法 4-2 [或子问题（Re-optimize）]，能够为原问题（P）更新上限 $\bar{Z}$。值得注意的是，对于约束非常紧的实例，启发式方法可能无法在某些迭代中返回可行的解决方案。另一个值得注意的地方是，这些启发式方法可以直接用于扩展模型，但是非线性功耗和网络拥塞等，可能导致上下界更加宽松。

### 4.4.3　使用加强割的拉格朗日松弛算法

前述的拉格朗日松弛算法可以应用到不同模型及其拓展中去。此节将深度挖掘 4.3.2 节中终端延迟成本项的结构性质，并以此为依据，为基础模型提出改良的拉格朗日松弛算法。将问题（P）中的容量限制约束和资源比例约束全部放松，以便拆分原问题。此时，资源供给决策就可以有显式解，超模成本方程将转化为次模方程。剩余的子问题仍可以转化为 MISOCP 来直接使用商业软件求解，但可以利用其结构性质（即次模性）加入加强割平面，加快求解速度。值得注意的是，次模性在 4.3 节中的其他的拓展模型可能并不成立，但是上一小节的拉格朗日算法还是可以解决这些拓展模型的。

对于问题（P），放松资源比例约束式(4-1f)和电能消耗约束式(4-1e)可以得到：

$$\min_{\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z}}\ (4\text{-}1a) + \sum_{j\in\mathcal{J}} \lambda_j \left( \sum_{k\in\mathcal{K}} w_k z_{jk} - p_j \right) + \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}} \sum_{l\in\mathcal{K}} \zeta_{jkl} \left( z_{jk} - \bar{r}_{kl} z_{jl} \right)$$

s.t. 约束式(4-1b)～ 式(4-1d)，式(4-1j)，式(4-1k)

其中 $(\lambda, \zeta)$ 分别是约束式(4-1e) 和式(4-1f)对应的拉格朗日乘子。给定需求分配决策 $y$ 后，对于数据中心 $j$ 的资源 $k$，优化拉格朗日子问题中的 $z_{jk}$ 带来了如下问题：

$$\min_{z_{jk}\geqslant 0} \phi_{jk} z_{jk} + \frac{\sum_{i\in\mathcal{I}} \tau_i d_i u_{ik} y_{ij}}{z_{jk} - \sum_i d_i u_{ik} y_{ij}}$$

其中 $\phi_{jk} = (\alpha c_j + \lambda_j) w_k + \sum_{l\in\mathcal{K}} (\zeta_{jkl} - \bar{r}_{lk}\zeta_{jlk})$。由拉格朗日对偶性质可知 $\phi_{jk} > 0$，因此最优资源供给决策为：

$$z^*_{jk} = \sum_i d_i u_{ik} y_{ij} + \sqrt{\sum_{i\in\mathcal{I}} \tau_i d_i u_{ik} y_{ij} / \phi_{jk}}$$

最优资源供给决策下的最优成本见式 (4-18)：

$$\phi_{jk}\sum_{i} d_i u_{ik} y_{ij} + 2\sqrt{\phi_{jk}\sum_{i\in\mathcal{I}} \tau_i d_i u_{ik} y_{ij}}. \tag{4-18}$$

令 $g_{jk}(y_{\cdot j}) = \sqrt{\phi_{jk}\sum_{i\in\mathcal{I}} \tau_i d_i u_{ik} y_{ij}}$。代入最优资源供给决策 $z_{jk}$，拉格朗日子问题变为式 (4-19)：

$$\begin{aligned}
\textbf{(P-LC)}\quad & \min_{\boldsymbol{x},\ \boldsymbol{y}} \sum_j f_j x_j + \sum_i \sum_j \left( d_i t_{ij} + \sum_k \phi_{jk} d_i u_{ik} \right) y_{ij} + \\
& 2\sum_j \sum_k v_{jk} - \sum_j \lambda_j p_j \\
\text{s.t.}\quad & g_{jk}(y_{\cdot j}) \leqslant v_{jk}, \quad \forall j \in \mathcal{J},\ k \in \mathcal{K} \\
& \text{约束式(4-1b)，式(4-1c)，式(4-1j)}
\end{aligned} \tag{4-19}$$

约束式(4-19)可转化为二阶锥约束。因此子问题（P-LC）也可以直接使用商业软件求解。但是商业软件中所使用的标准的分支定界法可以通过加入加强割平面来提高求解速度。其方法来源于 Atamtürk & Narayanan（2008）。定义多面体 $\mathcal{Q}_{jk}$ 来表示 $g_{jk}(\cdot)$ 的下凸包：

$$\mathcal{Q}_{jk} = \operatorname{conv}\{(\eta,\ t) \in \{0,\ 1\}^{|\mathcal{I}|} \times \mathbb{R} : g_{jk}(\eta) \leqslant t\}$$

另外，令：

$$\mathrm{EP}_{jk} = \left\{\pi \in \mathbb{R}^{|\mathcal{I}|} : \pi^{\mathrm{T}}\eta \leqslant g_{jk}(\eta),\ \ \forall \eta \in \{0,\ 1\}^{|\mathcal{I}|}\right\}$$

Atamtürk & Narayanan（2008）表明，线性不等式见式 (4-20)：

$$\pi^{\mathrm{T}}\eta \leqslant t \tag{4-20}$$

对于 $\mathcal{Q}_{jk}$ 是有效的当且仅当 $\pi \in \mathrm{EP}_{jk}$。因此将不成立的不等式(4-20)作为格外的割加入分支定界的过程中去以加快求解速度。展开来说，对于每个分式的解 $\hat{y}_{\cdot j} \in [0,\ 1]^{|\mathcal{I}|}$ and $\hat{v}_{jk} \geqslant 0$，定义式 (4-21)：

$$\hat{\pi}_{jk} = \arg\max\{\boldsymbol{\pi}^{\mathrm{T}}\hat{\boldsymbol{y}}_j,\ \boldsymbol{\pi} \in \mathrm{EP}_{jk}\} \tag{4-21}$$

如果 $(\hat{\boldsymbol{\pi}}_{jk})^{\mathrm{T}}\hat{\boldsymbol{y}}_j > \hat{v}_{jk}$，那么将下列割加入松弛问题（P-LC）中 [见式 (4-22)]：

$$(\hat{\pi}_{jk})^{\mathrm{T}} y_{\cdot j} \leqslant v_{jk} \tag{4-22}$$