



信息通信领域产学研合作特色期刊 十佳皖刊
第三届国家期刊奖百种重点期刊 中国科技核心期刊

ISSN 1009-6868
CN 34-1228/TN

中兴通讯技术

ZTE TECHNOLOGY JOURNAL

<http://tech.zte.com.cn>

2020年8月·第4期

专题：无线网络空中计算

1995

2G

3G

4G

5G

2020

25G

创刊 25 周年

纪念特刊

1995—2020



《中兴通讯技术》第8届编辑委员会成员名单

顾问 侯为贵(中兴通讯股份有限公司创始人) 钟义信(北京邮电大学教授) 陈锡生(南京邮电大学教授)

主任 陆建华(中国科学院院士)

副主任 李自学(中兴通讯股份有限公司董事长) 糜正琨(南京邮电大学教授)

编委(按姓名拼音排序)

- | | | | |
|-----|-----------------|-----|-------------------------|
| 陈建平 | 上海交通大学教授 | 唐雄燕 | 中国联通网络技术研究院首席科学家 |
| 陈前斌 | 重庆邮电大学教授、副校长 | 陶小峰 | 北京邮电大学教授 |
| 葛建华 | 西安电子科技大学教授 | 王文博 | 北京邮电大学教授、副校长 |
| 管海兵 | 上海交通大学教授 | 王文东 | 北京邮电大学教授 |
| 郭庆 | 哈尔滨工业大学教授 | 王喜瑜 | 中兴通讯股份有限公司执行副总裁 |
| 洪波 | 中兴发展股份有限公司总裁 | 王翔 | 中兴通讯股份有限公司高级副总裁 |
| 洪伟 | 东南大学教授 | 卫国 | 中国科学技术大学教授 |
| 黄宇红 | 中国移动研究院副院长 | 吴春明 | 浙江大学教授 |
| 纪越峰 | 北京邮电大学教授 | 邬贺铨 | 中国工程院院士 |
| 江涛 | 华中科技大学教授 | 肖甫 | 南京邮电大学教授 |
| 蒋林涛 | 中国信息通信研究院科技委主任 | 解冲锋 | 中国电信研究院教授级高工 |
| 李尔平 | 浙江大学教授 | 徐安士 | 北京大学教授 |
| 李红滨 | 北京大学教授 | 徐子阳 | 中兴通讯股份有限公司总裁 |
| 李厚强 | 中国科学技术大学教授 | 续合元 | 中国信息通信研究院副总工 |
| 李建东 | 合肥工业大学教授、副校长 | 薛向阳 | 复旦大学教授 |
| 李军 | 清华大学教授 | 薛一波 | 清华大学教授 |
| 李乐民 | 中国工程院院士 | 杨义先 | 北京邮电大学教授 |
| 李融林 | 华南理工大学教授 | 杨震 | 南京邮电大学教授、原校长 |
| 李少谦 | 电子科技大学教授 | 叶茂 | 电子科技大学教授 |
| 李自学 | 中兴通讯股份有限公司董事长 | 易芝玲 | 中国移动研究院首席科学家 |
| 林晓东 | 中兴通讯股份有限公司副总裁 | 张宏科 | 北京交通大学教授 |
| 刘健 | 中兴通讯股份有限公司高级副总裁 | 张平 | 中国工程院院士 |
| 刘建伟 | 北京航空航天大学教授 | 张卫 | 复旦大学教授 |
| 陆建华 | 中国科学院院士 | 张云勇 | 中国联通集团产品中心总经理 |
| 马建国 | 广东工业大学教授 | 赵慧玲 | 工业和信息化部通信科技委信息通信网络专家组组长 |
| 孟洛明 | 北京邮电大学教授 | 郑纬民 | 中国工程院院士 |
| 糜正琨 | 南京邮电大学教授 | 钟章队 | 北京交通大学教授 |
| 任品毅 | 西安交通大学教授 | 周亮 | 南京邮电大学教授 |
| 石光明 | 西安电子科技大学教授、副校长 | 朱近康 | 中国科学技术大学教授 |
| 孙知信 | 南京邮电大学教授 | 祝宁华 | 中国科学院半导体研究所研究员 |
| 谈振辉 | 北京交通大学教授、原校长 | | |

继往开来 续写新篇章

——《中兴通讯技术》创刊 25 周年纪念

◎ 文 / 编辑部

《中兴通讯技术》自 1995 年 7 月创刊，已经走过了 25 年的发展历程。25 个寒来暑往，25 个春华秋实，刊物忠实地记录了信息通信技术从 2G 到 5G 快速发展的风起云涌，见证了民族通信产业强势崛起的日新月异。25 年来，4 000 多位作者通过本刊跟广大读者分享了 2 000 多篇精彩论文，成就了 153 期厚重专题，助力企业实现“科技公益”的初心。

回顾 25 年办刊历程，我们在创新中发展，在发展中探索，完成了 5 个“五年计划”，走出了一条具有自身特色的办刊之路。第 1 个“五年”，汇聚优质办刊资源，高起点创办《中兴通讯技术》，成为企业新“名片”；第 2 个“五年”，伴随中兴通讯国际化步伐，创办《ZTE Communications》，刊物发往 140 多个国家；第 3 个“五年”，抓住 3G 带来的产业发展机遇，以刊为媒，成立产学研合作论坛；第 4 个“五年”，产学研与办刊工作融合发展，全面拓展刊物企业价值和社会价值；第 5 个“五年”，走“科技公益”之路，办“产学研特色期刊”，利用新媒体寻求刊物更加广泛和及时的传播，全面提升刊物竞争力和影响力。

刚刚过去的 5 年，是从 4G 迈入 5G 的 5 年，也是中国信息通信业与发达国家“并肩前行”到实现“超越发展”的 5 年。这期间，《中兴通讯技术》出版了 30 期技术专题，涵盖了当期信息通信领域的前沿技术和热点技术。回顾这些内容，可以将其聚类为以下 6 大专题：大数据与人工智能技术、云计算与边缘计算技术、物联网与

区块链技术、硅基光电子技术、网络重构与安全技术、5G 通信技术及其应用。这些我们耳熟能详的新技术，代表的是科技发展和时代进步，是无数科技工作者智慧和汗水的结晶。

过去的 5 年，是中兴通讯遭受挫折并再度扬帆起航的 5 年。2016 年和 2018 年公司两度面临困难和危机，业务和品牌都受到了极大影响。新一届管理层带领 8 万名员工鼓足干劲再出发，坚持技术领先，引领 5G 产品创新和行业落地；坚持全球化战略，积极把握政企和 5G 终端等新机遇；坚守合规经营，强化内控治理，防范系统性风险，确保经营可持续发展。目前，公司正以崭新的面貌，步入战略发展期。

过去的 5 年，对于中兴通讯技术杂志社来说，是不平凡的 5 年，更是砥砺奋进的 5 年。

首先是经历了编委会换届。侯为贵、钟义信、陈锡生、赵厚麟、乐光新、程时端等一批前辈因年事已高或公务繁忙，不再担任编委，继而由陆建华院士、高文院士接任中英文两刊编委会主任，并新增 22 位中青年编委。目前，两刊编委已达 105 人，他们是刊物健康发展的支撑和保障。

其次是完成了办刊资质的变更。在国家要求非时政类期刊必须转企改制的政策下，杂志社从 2012 年开始，进行了长达 6 年的转企改制工作。2018 年 12 月 25 日，国家新闻出版署正式批复同意两刊变更主管、主办和出版单位。主管单位由安徽省科技厅变更为安徽出版集团有限责任公司，主办单位由安徽省科学技术情

报研究所、中兴通讯股份有限公司变更为时代出版传媒股份有限公司、深圳航天广宇工业有限公司，出版单位由中兴通讯技术杂志社变更为安徽科学技术出版社。上述办刊资质的变更不仅使办刊主体完全符合国家政策要求，而且使得刊物的未来发展可以获取更多的出版资源。

再就是网络出版和新媒体传播迈出了坚实的步伐。2016年杂志社门户网站正式上线，2017年英文刊 ScholarOne 采编平台正式上线，2019年英文刊独立网站正式上线，2020年中文刊中国知网采编平台全面优化，2020年刊物排版设计开始采用方正学术出版云服务平台。这些举措帮助我们实现了平台采编、网络排版、免费获取和优先出版。在新媒体传播方面，我们做了多方面尝试，最终认为电子邮件推送是最适合本刊的精准传播方式，因为传播量和传播价值决定了一切。

过去的5年，更是我们坚定办刊方向，明确办刊道路的5年。企业为什么会办宣传刊物？因为要宣传其产品和品牌；企业为什么会办公开学术刊物？显然不为宣传自身产品，定位也绝不是宣传自身品牌。25年的总结和思考，终于让我们理解了中兴通讯企业办刊的初衷——为了科技公益，也就是传播知识，促进产业发展，回馈社会。这是一个崇高的使命，是一个民族高科技通信企业应有的担当！全国有5000多种科技期刊，其中信息通信类期刊有153种。如何发挥高科技企业的办刊优势，避免同质化竞争？实践是最好的导师，10年的产学研合作给我们指明了方向——走产学研合作的特色办刊之路，也就是借助中兴通讯产学研合作论坛这个平台，走产学研和办刊融合发展的道路。

一切过往，皆为序章，世界正经历着百年未有之大变局。信息通信技术加速了社会的发展，并深刻地改变着人们的生活。4G当下，5G应用已先声夺人，势如万马奔腾；6G概念，“犹抱琵琶半遮面”，让人们联想无限。学术界“破四唯”

优化评价机制，反造假严查学术不端；期刊界改革进入“深水区”，西方出版集团及其代理人长期掠夺出版资源的局面正在悄然改变，具有中国特色的期刊评价体系呼之欲出。中国期刊正在步入具有自身特色的发展道路。2020年5月28日，新闻出版署颁布了《报纸期刊质量管理规定》，明确了《期刊编校差错率计算方法》和《期刊出版形式差错数计算方法》，在内容、编校、出版形式、印刷4个方面推行“一票否决”制。可以预见，期刊的改革将会加速。

未来有无限的变化和可能，而不变的是《中兴通讯技术》的办刊宗旨：以人为本，荟萃通信技术领域精英；迎接挑战，把握世界通信技术动态；立即行动，求解通信发展疑难课题；励精图治，促进民族信息产业崛起。为了刊物发展的百年大计，我们将牢记科技公益的使命，不忘服务社会的初心，坚持走产学研结合的特色办刊之路。

千里之行，始于足下。下一个“五年”将是杂志社“整合资源，协同创新”的5年。当前我们已经拥有较好的资源平台，接下来需要汇聚更多资源，特别是用好这些资源，实现合作共赢和价值最大化，这才是我们刊物存在的价值，也是编辑人新时期面临的新任务。

刊物是桥梁，是纽带，维系着策划人、撰稿人、审稿人、编者、读者等多方的互信关系。刊物价值及其影响力决定了它汇聚资源的能力。因此，刊物的内涵必须像桥梁般可靠，刊物的外延需要像丝带般柔美。办有观点的刊物，打造有温度的平台，是我们不懈的追求。我们有信心在各级期刊管理部门的领导下，在编委的关爱和呵护下，在广大作者的信任和支持下，与改革同步，与时代同行，抓质量，出精品，亮特色。

25岁的《中兴通讯技术》，风华正茂，意气风发；25岁的《中兴通讯技术》，油墨飘香，待谱华章。

热烈祝贺

《中兴通讯技术》创刊 25 周年!



李自学

中兴通讯股份有限公司董事长

贺《中兴通讯技术》杂志创刊二十五周年

企业办刊搭建平台

产研合作共赢未来

公元 2020 年 李自学



邬贺铨

中国工程院院士

二十五载峥嵘岁月
产业报国矢志不渝
集智聚能越做越好
创新引领更上层楼

——祝贺《中兴通讯技术》
创刊25周年

邬贺铨
2020年7月17日



李乐民

中国工程院院士

引领信息与通信科技
的创新发展

祝贺《中兴通讯技术》创刊25周年

李乐民 2020年7月5日

通信技术的发展多姿多彩，
核心灵魂是智能化

—— 庆贺《中兴通讯技术》
创刊 25 周年

钟义信



钟义信

发展中国家工程院院士

高科技企业办名利
产学研合作结硕果

—— 祝贺《中兴通讯技术》杂志
创刊 25 周年

高文



高文

中国工程院院士



郑纬民

中国工程院院士

热烈祝贺《中兴通讯技术》创刊25周年。25年耕耘，成绩卓著，广传知识，产学研结合，为我国通讯事业的发展作出了很大贡献。祝期刊越办越好。

清华大学
郑纬民
2020年7月2日



张平

中国工程院院士

值此《中兴通讯技术》创刊25年之际，谨向为刊物做出贡献的人们致以崇高的敬意。祝《中兴通讯技术》杂志越办越好，与我国的信息技术一起茁壮成长。

张平
2020.7.4

传播通信知识
促进校企合作
吸引高端人才
服务产业未来

——祝贺《中兴通讯技术》杂志
创刊 25 周年

庄卫华



庄卫华

加拿大工程院院士
加拿大皇家科学院院士

助推通信理论突破变革世界
传扬网络技术创新造福人类

祝贺《中兴通讯技术》创刊 25 周年

任福继

2020年7月2日



任福继

日本工程院院士

中兴通讯技术杂志社办刊25周年大事记

1995— 2000 年 开拓创新的五年

- 1995年6月,《中兴新通讯》编辑部成立,编委会组建
- 1995年7月,《中兴新通讯》正式创办,获省级刊号,季刊出版
- 1997年1月,《中兴新通讯》改为双月刊
- 2000年7月,《中兴新通讯》获正式刊号,国内外公开发行,同时更名为《中兴通讯技术》
- 2000年10月,《中兴通讯技术》被中文科技期刊数据库收录

2001— 2005 年 快速发展的五年

- 2001年8月,《中兴通讯技术》获安徽省优秀科技期刊二等奖
- 2002年9月,《中兴通讯技术》获第3届华东地区优秀期刊奖
- 2002年12月,《中兴通讯技术》被中国核心期刊(遴选)数据库收录
- 2003年6月,《ZTE Communications》创刊,季刊出版,中兴通讯技术杂志社成立
- 2003年7月,《中兴通讯技术》入选为中国学术期刊综合评价数据库(CAJCED)统计源期刊
- 2003年7月,《中兴通讯技术》被中国期刊全文数据库(CJFD)全文收录
- 2003年12月,《中兴通讯技术》荣获首届《CAJ-CD规范》执行优秀期刊奖
- 2004年1月,《中兴通讯技术》被收录为中国科技论文统计源期刊(中国科技核心期刊)
- 2005年1月,《中兴通讯技术》荣获第3届全国期刊奖百种重点期刊称号
- 2005年10月,《ZTE Communications》获正式刊号,国内外公开发行

2006— 2010 年 国际化发展的五年

- 2006年6月,《ZTE Communications》被中国期刊全文数据库(CJFD)全文收录
- 2008年7月,《ZTE Communications》被美国《乌利希期刊指南》收录
- 2009年3月,《ZTE Communications》被美国《剑桥科学文摘(工程技术)》(CSA(T))收录
- 2009年5月,《ZTE Communications》被波兰《哥白尼索引》(Index Copernicus)收录
- 2009年11月,《中兴通讯技术》获安徽省优秀期刊奖
- 2009年11月,《中兴通讯技术》获第4届华东地区优秀期刊奖
- 2010年5月,《ZTE Communications》被中国核心期刊(遴选)数据库收录

2011— 2015 年 产学研合作的五年

- 2011 年 1 月,《ZTE Communications》被中文科技期刊数据库收录
- 2011 年 3 月,《ZTE Communications》组建编委会,刊物开始独立组稿、独立运作
- 2011 年 12 月,《中兴通讯技术》获工业和信息化部 2010—2011 年度优秀科技期刊奖
- 2012 年 7 月,建立编委任期制,并发布编委聘任制度文件
- 2012 年 12 月,《中兴通讯技术》获 2012 年华东地区优秀期刊奖
- 2012 年 12 月,《ZTE Communications》被英国 INSPEC 数据库收录
- 2013 年 2 月,中英文两刊引入 DOI 并启用数字优先出版
- 2013 年 6 月,《ZTE Communications》被挪威 NSD 数据库收录
- 2014 年 8 月,产学研项目的创新成果成为《ZTE Communications》稿源
- 2015 年 1 月,调整中英文两刊栏目,体现刊物观点性和原创性
- 2015 年 2 月,明确走办产学研合作特色期刊之路
- 2015 年 3 月,《中兴通讯技术》被评为 RCCSE (中国学术期刊评价研究中心) 中国核心学术期刊 (A)
- 2015 年 4 月,明确数字化出版发展路径,筹建刊物门户网站
- 2015 年 6 月,出版纪念创刊 20 周年特刊及画册
- 2015 年 8 月,完成编委会换届,成立第 7 届编委会

2016— 2020 年 继往开来的五年

- 2016 年 6 月,《中兴通讯技术》独立网站 tech.zte.com.cn 上线
- 2016 年 10 月,《中兴通讯技术》获首届安徽省科技期刊编校质量优秀奖
- 2016 年 12 月,《ZTE Communications》启用 ScholarOne 稿件管理平台
- 2017 年 2 月,《中兴通讯技术》获 2015—2016 年度安徽省优秀期刊奖
- 2017 年 3 月,《ZTE Communications》被俄罗斯《文摘杂志》(Abstract Journal) 收录
- 2018 年 8 月,完成编委会换届,成立第 8 届编委会
- 2018 年 9 月,《中兴通讯技术》获十佳皖版期刊奖
- 2018 年 10 月,《ZTE Communications》被评为 2018 中国国际影响力优秀学术期刊
- 2018 年 12 月,《中兴通讯技术》和《ZTE Communications》完成主管、主办和出版单位变更
- 2019 年 2 月,《中兴通讯技术》和《ZTE Communications》封面和文章首页添加二维码
- 2019 年 5 月,《中兴通讯技术》被日本科学技术振兴机构数据库 (JST) 收录
- 2019 年 6 月,《中兴通讯技术》入选庆祝中华人民共和国成立 70 周年精品期刊展
- 2019 年 9 月,杂志社应邀在第 15 届中国科技期刊发展论坛介绍产学研合作特色办刊经验
- 2019 年 10 月,《ZTE Communications》被评为 2019 中国国际影响力优秀学术期刊
- 2019 年 10 月,杂志社应邀在 2019 中国学术期刊未来论坛上介绍产学研合作特色办刊经验
- 2019 年 11 月,《ZTE Communications》独立网站 tech-en.zte.com.cn 上线
- 2019 年 12 月,《中兴通讯技术》被评为 2017—2018 年度安徽省优秀期刊
- 2020 年 3 月,刊物排版设计开始采用方正学术出版云服务平台
- 2020 年 3 月,开展“知识服务”活动,征集到 5 500 多个目标读者邮箱
- 2020 年 4 月,开展“严抓‘三审三校’,提升编校质量”专项活动
- 2020 年 8 月,出版创刊 25 周年纪念特刊

ZTE中兴

5G

热烈祝贺
中兴通讯股份有限公司
成立 35 周年

1985—2020

目次

中兴通讯技术 (ZTE TECHNOLOGY JOURNAL)
总第 153 期 第 26 卷 第 4 期 2020 年 8 月

专题：无线网络空中计算

移动边缘计算中的资源管理 **02**
游昌盛

大规模移动边缘计算网络：空间建模及
计算吞吐量优化 **06**
韩凯峰，胡昌军，刘铁志

基于神经网络计算的无线容量高实时预测 **13**
赖昱辰，钟祎，王建峰

基于空中计算的无线群智感知 **18**
李晓阳，贡毅

面向高效通信边缘学习网络的
通信计算一体化设计 **23**
朱光旭，李航

面向边缘智能的空中计算 **31**
曹晓雯，莫小鹏，许杰

专家论坛（创刊 25 周年特约稿）

38 万物互联，任重道远
李少谦

40 网络管理自动化中闭环形成的概念
孟洛明

43 通信产业发展回顾与展望
张云勇

46 知识 + 数据驱动学习：未来网络智能的基础
朱近康

50 针对 5G/B5G 的大规模 MIMO 系统射频前端设计
马建国

58 无线物理层认证技术：昨天、今天和明天
任品毅，徐东阳

企业视界

67 确定性网络技术及应用场景研究
魏月华，喻敬海，罗鉴

2020 年第 1—6 期专题计划及策划人

1. 蜂窝车联网产业与技术

中国移动通信研究院首席科学家 易芝玲
中国移动通信研究院技术经理 潘成康

2. 智能化通信应用芯片技术

中国科学院半导体研究所研究员 祝宁华
中国科学院半导体研究所研究员 李明

3. 5G 核心网技术与挑战

工业和信息化部通信科技委
信息通信网络专家组组长 赵慧玲

4. 无线网络空中计算

中国科学技术大学教授 卫国
中国科学技术大学副研究员 陈力

5. 网络人工智能技术

电子科技大学教授 虞红芳

6. 工业互联网技术与应用

中国信息通信研究院副总工 续合元

CONTENTS

ZTE TECHNOLOGY JOURNAL Vol. 26 No. 4 Aug. 2020

Special Topic:

Over-the-Air Computation for Wireless Network

Resource Management in Mobile Edge Computing **02**
YOU Changsheng

Large-Scale Mobile Edge Computing Network:
Spatial Modeling and Computation Throughput
Optimization **06**
HAN Kaifeng, HU Changjun, LIU Tiezhi

High Real-Time Capacity Prediction Based on
Neural Network Evaluation **13**
LAI Yuchen, ZHONG Yi, WANG Jianfen

Over-the-Air Computation
Based Wireless Crowd Sensing **18**
LI Xiaoyang, GONG Yi

Integrating Communication and Computation for
Communication-Efficient Edge Learning over
Wireless Networks **23**
ZHU Guangxu, LI Hang

Over-the-Air Computation for Edge Intelligence **31**
CAO Xiaowen, MO Xiaopeng, XU Jie

Expert Forum

38 Interconnection of Everything Has a Long Way to Go
LI Shaoqian

40 Closed Loop in Autonomous Network Management
MENG Luoming

43 Review and Prospect of Communications Industry
Development
ZHANG Yunyong

46 Knowledge-and-Data Driven Learning: Foundation of
Future Network Intelligence
ZHU Jinkang

50 RF Front-End Designs of MIMO Systems for
5G and Beyond
MA Jianguo

58 Wireless Physical Layer Authentication Technology:
Yesterday, Today, and Tomorrow
REN Pinyi, XU Dongyang

Enterprise View

67 Deterministic Networking Technology and Scenarios
WEI Yuehua, YU Jinghai, LUO Jian

期刊基本参数: CN 34-1228/TN*1995*b*16*72*zh*P* ¥ 20.00*15000*13*2020-08

敬告读者

本刊享有所发表文章的版权, 包括英文版、电子版、网络版和优先数字出版版权, 所支付的稿酬已经包含上述各版本的费用。未经本刊许可, 不得以任何形式全文转载本刊内容; 如部分引用本刊内容, 须注明该内容出自本刊。



无线网络空中计算专题导读

专题策划人 卫国



中国科学技术大学教授，曾任国家“863”计划通信技术主题专家组成员、中国第三代移动通信系统研究开发项目总体组成员、国家“863”计划 B3G 移动通信重大项目总体组成员、“新一代宽带无线移动通信网”国家科技重大专项总体专家组成员；主要从事无线通信技术、移动通信网络、信号处理等方面的研究；获国家科技进步二等奖 1 项；发表论文 100 余篇，拥有国家发明专利 10 余项。

专题策划人 陈力



中国科学技术大学信息技术学院副研究员；主要从事无线通信、通信计算融合、通信感知融合等相关研究；负责国家自然科学基金、国家重大专项子课题等多个研究项目；曾获中国科学院院长优秀奖；发表 SCI 期刊论文 30 余篇，拥有国家发明专利 10 余项。

5G 移动通信网络的海量接入能力和低时延传输特征，一方面为物联网开辟了一片新的广阔天地，另一方面也释放出强烈的信号：未来移动通信网不再只是为了满足人们的通信需求，而是朝着更为广泛的物与物之间的数据连接演进。

无线节点的海量增长，带来的是高达百亿的内连接数量与每年数泽字节的数据总生成量。无线网络由于其资源限制，无论采用怎样的接入技术，汇聚如此海量的数据都成为一件非常困难的事。此外，对于海量数据的计算处理，物联网节点也同样面临着巨大挑战。这就催生了无线空中计算这个新的研究方向。

机器学习通过大数据的分析与处理，使得越来越多的领域智能化，并产生了更多新功能的应用。无线网络天然具备海量的数据，而利用机器学习锤炼这些数据，打造无线网络的新功能是一种值得探索的可能。与传统机器学习的架构不同，无线网络通常是分布式的层次化架构，并受制于节点性能与无线链路状态。这将为无线网络智能化带来全新挑战。

无线网络空中计算的核心问题，一方面在于如何解决海量数据收集与大规模计算带来的传输时延与计算时延，另

一方面在于如何设计适合于无线网络的智能计算框架。针对这些挑战性的问题，本期专题提供了一个讨论的平台。《移动边缘计算中的资源管理》与《大规模移动边缘计算网络：空间建模及计算吞吐量优化》从建模、网络规模与资源管理等方面为移动边缘计算架构提供了重要的设计指南。该网络架构为无线网络创造无处不在的快速计算环境，将计算任务迁移到边缘端来降低计算时延与数据收集量。《基于神经网络计算的无线容量高实时预测》指出充分利用海量节点提供的信道状态信息，并基于神经网络，能够为无线网络提供无线容量实时预测的新功能。《基于空中计算的无线群智感知》引入了空中计算技术来实现网络的智能感知功能。该技术利用无线多址接入信道的信号叠加特性，能在信号传输的同时完成目标函数的计算，从而降低传输开销。《面向高效通信边缘学习网络的通信计算一体化设计》与《面向边缘智能的空中计算》提出机器学习、边缘计算与空中计算的结合，可以为无线网络提供新功能，并降低传输与计算带来的时延。

上述工作基本上反映出在空中计算这一方向上中国研究者的主要成果与学术观点，从不同侧面为该领域的研究展示了多种可能。希望能对无线网络新技术的研究发展起到一定的推动作用。

移动边缘计算中的 资源管理

Resource Management in Mobile Edge Computing



游昌盛 / YOU Changsheng

(新加坡国立大学, 新加坡 117576)
(National University of Singapore, Singapore 117576, Singapore)

摘要: 通过对移动边缘计算 (MEC) 网络的基本原理、应用场景, 以及通信和计算的研究模型的阐述, 提出了针对单用户和多用户 MEC 系统的绿色节能频谱和计算资源综合管理方案。通过分析当前 MEC 技术的局限和挑战, 认为 MEC 和人工智能技术的有机结合, 能够有效提高未来网络的计算性能。

关键词: 移动边缘计算; 无线通信; 资源管理

Abstract: For mobile edge computing (MEC), the principles, use cases, and its communication-and-computation modelling are introduced. Then, a set of energy-efficient joint radio-and-computational resource management is proposed for single-user and multi-user MEC systems. Finally, in view of existing limitations and challenges in MEC systems, an outlook on its seamless integration with artificial intelligence (AI) for improving the computing performance of future networks is provided.

Keywords: mobile edge computing; wireless communication; resource management

DOI: 10.12142/ZTETJ.202004002
网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20200714.0933.002.html>

网络出版日期: 2020-07-14
收稿日期: 2020-06-04

1 研究背景

随着 5G 技术和物联网 (IoT) 的发展, 无线设备的数目呈指数级增长, 物联网应用场景也越来越多样化。这其中包括大量计算密集型和时延敏感型的应用, 如虚拟现实、在线游戏等, 这类应用需要强大的计算能力支持来满足超低时延的要求。为了满足这种需求, 近年来, 传统的云计算网络架构正悄然向移动边缘计算 (MEC) 网络发生转变: 原本位于核心网云数据中心的计算服务和功能正在往网络边缘下沉, 通过离用户更近的基站和无线接入点向用户提供无处不在的计算、

存储、通信等服务, 从而有效降低用户的计算时延和能耗, 并大大提高整个网络的资源利用率^[1-4]。

如图 1 所示, 在 MEC 系统中, 用户可将本地计算密集型的任务卸载 (迁移) 到 MEC 服务器中, 让它代为计算并向用户反馈计算结果。与传统的基于数据中心的移动云计算 (MCC) 相比, MEC 拥有如下几方面明显优势:

1) 更低的时延: 由于边缘云离用户更近且计算数据迁移到云的过程中不涉及在核心网中的数据传输, MEC 可以大大降低 MCC 中的数据传播和核心网回程链路时延。另一方面, 通过广泛部署 MEC 服务器, 每台 MEC 服

务器仅需向周边少量用户提供计算服务, 从而达到较低的计算时延。因此, 相较于 MCC 所需要的 100 ms 量级时延, MEC 可满足 1~10 ms 量级的超低时延要求。

2) 更低的能耗: MEC 用户可选择将高能耗型的计算任务迁移到边缘云中, 从而避免本地计算带来的巨大能耗。另一方面, 由于离 MEC 服务器更近, MEC 用户可以大大降低计算数据传输中的能量消耗。

3) 更优的情境感知: 利用近距离优势, MEC 服务器可以通过用户的定位信息等更加准确地预测和判断用户的计算行为和需求, 从而提供更及



▲图1 MEC系统架构与应用

时有效的计算和存储服务。

4) 更强的安全保护: 和 MCC 相比, MEC 服务器的用户数目更少, 且用户数据信息不需要经过复杂的核心网到达数据中心。这样可以有效缓解数据在多跳网络传输中的信息泄漏问题。

2 研究模型

为了研究 MEC 系统的计算性能, 我们首先介绍 MEC 的基本研究模型。

1) 计算任务模型: 总的来说, MEC 的计算卸载模型包括全部卸载和部分卸载。其中, 全部卸载计算模型适用于数据不可分割的高集成计算任务, 它要求用户只能选择全部本地计算或者全部卸载到 MEC 服务器。这类计算任务的关键参数包括: 计算数据量 (比特数)、计算强度 (每比特数据需要的中央处理器时钟数), 以及计算时延要求。这些参数与具体的计算任务相关, 可以通过对计算任务的

剖析和建模得出。另一方面, 部分卸载模型适用于两类计算任务: 一类是数据可任意分割的计算 (如数据压缩等); 另一类是包含多个子任务的计算, 不同任务间往往具有一定的运算顺序和联系, 如图 2 所示。与全部卸载模型相比, 部分卸载模型有更多的设计自由度, 可以更有效地卸载部分数据或子任务来减少用户的计算时延和能耗。

2) 计算时延和能耗模型: 对于用户的本地计算, 计算时延与计算所需的中央处理器 (CPU) 时钟数成正比, 与 CPU 主频成反比; 因此, 我们可以通过提高 CPU 的主频来降低本地计算时延, 但这样做同时也会增加本地计算的能耗。本地计算的能耗主要来自于 CPU 的功耗, 而 CPU 的功耗又与 CPU 主频的平方成正比; 因此, CPU 主频越高, 本地计算能耗越高且增长越快。对于 MEC 服务器 (或边缘

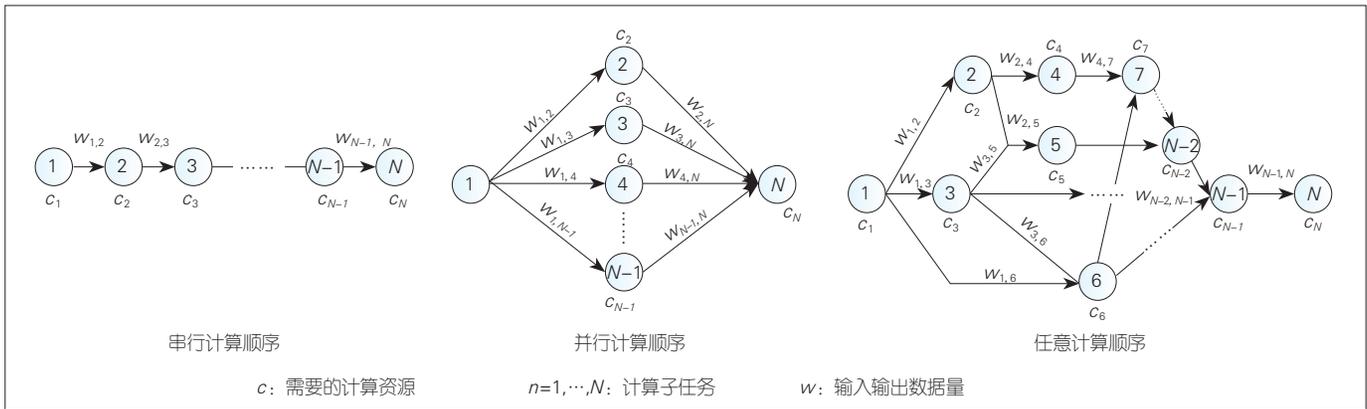
云) 来说, 它的计算时延包括 CPU 的运算时延和多计算任务下的队列时延, 能耗包括 CPU 的计算能耗以及服务器的开机运行能耗。

3) 无线数据传输模型: 5G 通信的各种技术, 如毫米波通信、非正交复用接入多址等都可以被有效用于提高计算卸载时数据传输的可达速率 (通过香农公式建模), 从而降低计算数据的传输时延。同时, 用户还可以利用设备到设备 (D2D) 的通信技术来实现低时延的用户间数据传输和计算卸载^[4]。

3 MEC 系统中的综合资源管理

无线和计算资源的综合管理是 MEC 系统设计中的重要组成部分。针对不同的 MEC 系统设置, 我们需要解决不同的综合资源管理问题。

首先考虑单用户情况下基于全部卸载计算模型的 MEC 资源管理。其中, 最重要的设计问题是如何做卸载决策,



▲图2 移动边缘计算任务不同的计算顺序

即是否进行计算卸载和如何设计卸载策略。为了研究这个问题,文献[6]提出了一个新型的无线供能下的MEC系统,并设计它最优的计算卸载方案。为了满足计算时延要求并最小化用户的计算耗能,我们分别优化了本地计算和计算完全卸载两种模式下的设计:对于本地计算,通过优化用户CPU的主频来降低计算能耗,同时满足计算耗能不大于获得的无线能量的条件;对于计算完全卸载,提出最优的时间分割方案,使用户能够先获取充足的能量然后进行数据传输和计算迁移,同时最大化用户的剩余能量。最后,基于本地计算和完全卸载两种模式的不同能耗,提出最优的本地计算/完全卸载的决策。这个工作后续被拓展到更加复杂的MEC系统,如基于能量收集的MEC系统^[7]和基于无线供能的多用户MEC系统^[8]。

对于多用户下的MEC系统而言,它的综合资源管理更加复杂。在文献[9]中,我们考虑部分卸载的计算模型并假设所有用户需要在相同的时间段内完成不同强度的计算任务。为了最小化所有用户的总计算能耗(包括每个用户的本地计算和卸载能耗),利用凸优化工具我们提出了一套最优的综合资源管理设计方案。具体来说,首先计算得到一个(计算)卸载优先

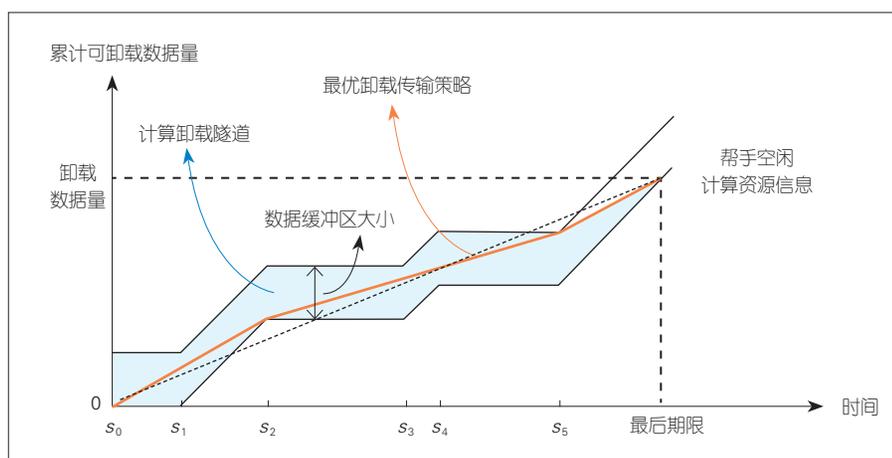
级函数,它与用户的信道增益和本地计算耗能成正比;因此,对于每个用户来说,他的无线信道越好或者本地耗能越大,用户的卸载优先级就越高。基于这个优先级函数,我们证明了最优的综合资源管理方案是一种基于门限的资源分配:对每个用户来说,如果他的卸载优先级函数值高于一定门限,他将选择把计算任务全部卸载到MEC服务器;反之,他将尽量在本地完成所有运算。与文献[9]中的集中式资源管理不同,文献[10]研究了分布式的计算资源分配。作者考虑了完全卸载的计算模型,并采用游戏理论来解决不同用户计算卸载与否的非确定性(NP)难问题。文献[10]中的研究证明,当用户受到的信号干扰强度低于一定门限时,他应该将计算任务卸载到云端。因为在这种情况下,无线传输可达速率较大,完成计算数据传输的能耗比本地计算更小。

文献[11]提出将基于边缘基站的MEC系统拓展到用户间计算卸载的MEC系统,从而有效降低边缘基站的计算和通信负载压力,并提高整个MEC系统的计算资源利用率。具体来说,主要利用用户周边的移动设备(如电脑等)的计算资源来支持用户的计算卸载。与基于边缘基站的MEC系统相比,用户周边的移动设备(简称帮手)

呈现出时有时无的空闲的计算资源。这是因为帮手只有在自己没有计算任务时,才能给周边的用户提供空闲的计算资源。利用这个特性,我们提出了一种基于帮手空闲计算资源的变速率计算卸载算法。该算法的核心在于首先在以横坐标为时间、纵坐标为累计可卸载数据量的坐标轴上构建一个“计算卸载隧道”,隧道的顶部和底部形状与帮手的缓存区大小和空闲计算资源的存量有关,具体如图3所示。为了最小化用户能耗,用户的计算卸载速率可以利用这个计算卸载隧道和几何方法得到。直观来看,如图3所示,这个方法就是在隧道的两端拉一条绷紧的线,不同线段的斜率反应了不同时间段内计算卸载的数据传输速率。这个方法可以被进一步拓展到多用户间的计算卸载场景。

4 MEC 未来工作展望

用人工智能算法设计MEC策略。当前的MEC策略设计主要有两种方法:一种是利用凸优化等优化理论来设计最优或次优的计算卸载策略,但是对于大规模MEC系统或优化问题本身是NP难的情况,用优化理论来设计MEC策略的方法可能需要很长的时间,这与MEC致力于缩短计算时延的初衷相违背;另一种方法是用启发式算法



▲图3 用户间计算卸载策略

来设计低复杂度的 MEC 策略，但这类方法往往缺乏一定的理论支撑，可能无法达到较好的 MEC 计算性能。为了解决这个问题，一个有效的方法是利用人工智能技术来实现快速高效的计算卸载策略设计。例如，我们可以将 MEC 策略优化问题转化为相应的深度学习问题：神经网络输入是用户的计算模型信息，神经网络的输出是计算卸载的策略。通过大量的计算卸载策略采样，我们可以训练出一个智能神经网络。这样一来，在实际的计算卸载决策中，我们只需将即时的计算模型信息输入到神经网络中，就可以快速得到一个有效的计算卸载策略方案。对于大规模 MEC 系统来说，基于凸优化理论的策略采样可能无法实现，但这时候我们可以利用小规模 MEC 系统进行神经网络训练，然后通过迁移学习的方法得到大规模 MEC 系统的策略采样。如何设计神经网络的训练是未来研究工作中的一个重要方向。

针对人工智能算法的 MEC 建模和设计。当前的 MEC 计算模型主要考虑普适性的计算，即计算数据量的大小与计算强度通常是一个固定的线性关系，但这个简单模型并不一定适用于具体的人工智能算法。例如，深度学习的计算复杂度除了和数据相关外，

还与神经网络的深度、每层网络的节点数、网络的类型（如卷积/自回归网络）等息息相关。因此，如何对具体人工智能算法进行计算模型建模是一个亟待探索和研究的重大问题。除此之外，当前的 MEC 策略设计主要关注计算时延和能耗，但这些性能指标并不完全是人工智能应用中最关心的问题。

5 结束语

MEC 将无线设备终端上计算密集型的运算任务迁移到边缘云中，从而有效降低用户的计算时延和能耗。在本文中，我们阐述了 MEC 系统的基本原理和模型，并提出如何紧密结合无线通信和计算机技术来设计不同 MEC 系统下的计算卸载策略。在无线网络智能化的关键时期，如何将 MEC 和人工智能技术有机结合来提升未来无线网络的性能需要继续探索。

参考文献

- [1] MAO Y Y, YOU C S, ZHANG J, et al. A survey on mobile edge computing: the communication perspective [J]. IEEE communications surveys & tutorials, 2017, 19(4): 2322–2358. DOI: 10.1109/comst.2017.2745201

- [2] YU W, LIANG F, HE X F, et al. A survey on the edge computing for the Internet of Things [J]. IEEE access, 2018, 6: 6900–6919. DOI: 10.1109/access.2017.2778504
- [3] 马洪源. 面向 5G 的边缘计算及部署思考 [J]. 中兴通讯技术, 2019, 25(3): 77–81. DOI: 10.12142/ZTETJ.201903011
- [4] QIN M, CHEN L, ZHAO N, et al. Power-constrained edge computing with maximum processing capacity for IoT networks [J]. IEEE Internet of things journal, 2019, 6(3): 4330–4343. DOI: 10.1109/ijot.2018.2875218
- [5] PU L J, CHEN X, XU J D, et al. D2D fogging: an energy-efficient and incentive-aware task offloading framework via network-assisted D2D collaboration [J]. IEEE journal on selected areas in communications, 2016, 34(12): 3887–3901. DOI: 10.1109/jsac.2016.2624118
- [6] YOU C S, HUANG K B, CHAE H. Energy efficient mobile cloud computing powered by wireless energy transfer [J]. IEEE journal on selected areas in communications, 2016, 34(5): 1757–1771. DOI: 10.1109/jsac.2016.2545382
- [7] MAO Y Y, ZHANG J, LETAIEF K B. Dynamic computation offloading for mobile-edge computing with energy harvesting devices [J]. IEEE journal on selected areas in communications, 2016, 34(12): 3590–3605. DOI: 10.1109/jsac.2016.2611964
- [8] BI S Z, ZHANG Y J. Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading [J]. IEEE transactions on wireless communications, 2018, 17(6): 4177–4190. DOI: 10.1109/twc.2018.2821664
- [9] YOU C S, HUANG K B, CHAE H, et al. Energy-efficient resource allocation for mobile-edge computation offloading [J]. IEEE transactions on wireless communications, 2017, 16(3): 1397–1411. DOI: 10.1109/twc.2016.2633522
- [10] CHEN X, JIAO L, LI W Z, et al. Efficient multi-user computation offloading for mobile-edge cloud computing [J]. ACM transactions on networking, 2016, 24(5): 2795–2808. DOI: 10.1109/tnet.2015.2487344
- [11] YOU C S, HUANG K B. Exploiting non-causal CPU-state information for energy-efficient mobile cooperative computing [J]. IEEE transactions on wireless communications, 2018, 17(6): 4104–4117. DOI: 10.1109/twc.2018.2820077

作者简介



游昌盛，新加坡国立大学博士后；主要研究方向为移动边缘计算、边缘学习、无人机通信、智能反射面通信等；2019 年获 IEEE 通信学会亚太地区优秀论文奖；已发表 SCI/EI 论文 20 余篇。

大规模移动边缘计算网络： 空间建模及计算吞吐量优化

Large-Scale Mobile Edge Computing Network:
Spatial Modeling and Computation Throughput Optimization



韩凯峰/HAN Kaifeng, 胡昌军/HU Changjun, 刘铁志/LIU Tiezhi

(中国信息通信研究院, 中国 北京 100191)

(China Academy of Information and Communications Technology, Beijing 100191, China)

摘要:提出并定义了大规模移动边缘计算(MEC)网络中的空间计算吞吐量这一性能指标,通过运用随机几何、凸优化等理论,对这一性能指标进行了分析和最优化设计。利用随机几何理论为大规模 MEC 网络建立空间模型,该模型涵盖边缘云及用户的空间随机分布、无线接入、计算任务卸载、边缘端并行计算等重要的网络特征。基于网络模型,首次对 MEC 网络空间计算吞吐量进行定义和分析,并通过优化设计 MEC 服务范围半径(r_0)以及用户计算卸载比例(ρ)这两个指标,来实现 MEC 网络空间吞吐量的最大化。所提供的严谨的理论分析、富有物理内涵的优化结果将为部署大规模 MEC 网络提供了极为重要的设计指南。

关键词:移动边缘计算;移动计算卸载;无线网络建模;随机几何;凸优化

Abstract: The attempt on defining, analyzing, and optimizing the spatial computation throughput in a large-scale wireless mobile edge computing (MEC) network is made. The analysis involves the interplay of theories of stochastic geometry and convex optimization. Specifically, the large-scale MEC network features of wireless access and edge-computing are first modeled, such as random nodes distribution, computation tasks offloading, parallel computing at edges, by using stochastic geometry. Based on the proposed model, the spatial computation throughput of the MEC network is defined, studied and maximized in terms of the radius of MEC service range (denoted by r_0) as well as offloading ratio of active mobile users (denoted by ρ) under the constraints of latency and energy. The optimal solutions of r_0 and ρ can be easily calculated via solving simple equations, and their closed-form results could be obtained in the extreme case. The tractable analysis and insightful results give useful design guidelines for MEC network planning and provisioning in a large-scale space.

Keywords: mobile edge computing; mobile computation offloading; wireless network modeling; stochastic geometry; convex optimization

DOI: 10.12142/ZTETJ.202004003

网络出版地址: <https://kns.cnki.net/KCMS/detail/34.1228.TN.20200715.1709.002.html>

网络出版日期: 2020-07-16

收稿日期: 2020-05-21

作为 5G 系统中的关键技术之一,移动边缘计算(MEC)可以利用部署在网络边缘的服务器为移动用户提供泛在、低时延的高质量计算服务^[1-2],例如多媒体云游戏、增强现实等。相比中心化的云计算,去中心化的 MEC 能够显著降低计算时延、移动

用户能耗以及网络传输复杂度^[3-4],适用于未来边缘智能网络^[5]、大规模物联网^[6]以及点对点网络^[7]等应用场景。

在 MEC 中,一个研究热点领域是设计高效能、低时延的移动计算卸载方式,即移动用户可将其计算任务卸载到 MEC 服务器中进行计算。文献

[8]提出了一个卸载计算策略,在给定计算时限要求下,通过联合设计频谱和计算分配资源来最小化移动用户的能耗。对于类似的优化目标,文献[9]给出了基于 Lyapunov 优化理论的动态优化结果。文献[10]则将移动计算卸载及资源分配设计问题拓展至

车联网场景。

虽然上述研究工作仅考虑包含一个或几个边缘云和用户的小规模 MEC 网络,就能设计出复杂但有效的计算卸载方案或策略,但是研究并设计并优化大规模 MEC 网络(包含无限多个边缘云和用户的 MEC 网络)中的无线通信和边缘计算的性能指标也同样重要。文献[11]首次提出了基于随机几何理论的大规模 MEC 网络模型,并对 MEC 网络的传输时延和计算时延进行了理论分析以及优化设计,为大规模 MEC 网络部署提供了重要设计指南;但该文献并未研究如何定义并最优化地设计 MEC 网络空间吞吐量的问题。

为此,我们提出并定义了大规模 MEC 网络中的空间计算吞吐量,并通过优化设计 MEC 服务范围半径以及用户计算卸载比例,来实现 MEC 网络空间吞吐量的最大化,以期部署大规模 MEC 网络提供参考。

1 系统模型及性能指标

考虑一个包含无限多个 MEC 服务器和移动用户的大规模 MEC 网络(如图1所示)。

1.1 MEC 网络空间分布

在二维空间内,假设 MEC 服务器和用户的位置都服从泊松点过程(PPP)分布,其中 MEC 服务器位置 $X \in \mathbb{R}^2$ 服从密度为 λ_s 的 PPP 分布 $\Omega = \{X\}$,用户位置 $Y \in \mathbb{R}^2$ 服从密度为 λ_u 的 PPP 分布 $\Phi = \{Y\}$ 。

1) 多用户接入模型

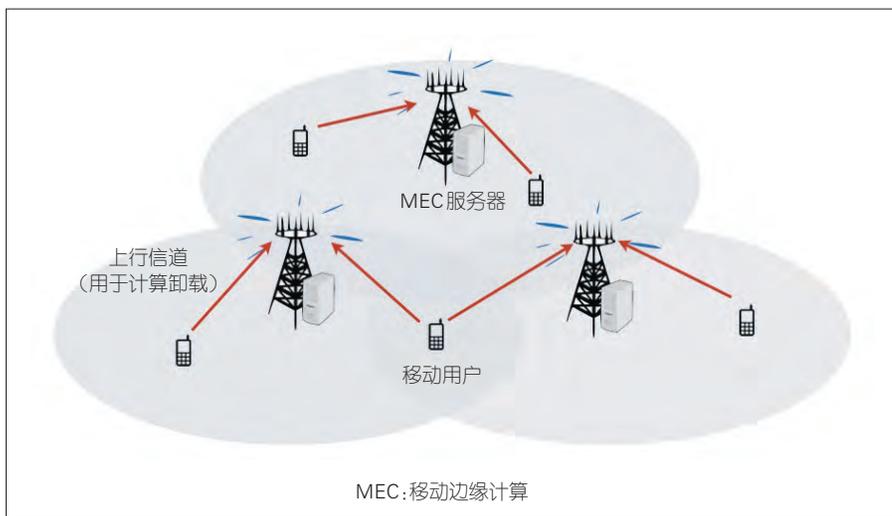
将总带宽资源分为 M 个正交的子信道 $\{1,2,\dots,M\}$,每个用户可以随机选择一个子信道将计算任务上传到 MEC 服务器进行计算(即计算卸载),选择同一子信道进行传输的多个用户将产生干扰。令时隙间隔为

T_s 秒,假设用户位置及信道状态在不同时隙相互独立,并要求每个用户需要在 T_s 秒内完成计算任务。

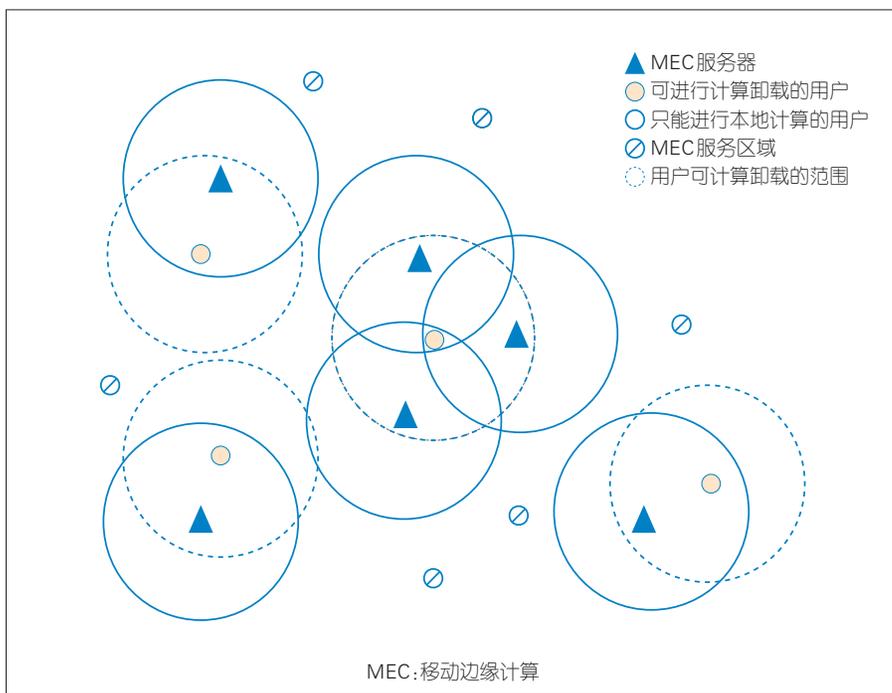
2) MEC 服务区域

图2展示了大规模 MEC 网络空间分布图。每个 MEC 服务器 X 均可形成一个半径为 r_0 的圆形 MEC 服务区域,表示为 $O = (X,r_0)$,其中 r_0 为常量。假设只有被 MEC 服务区域覆盖的用户可以选择计算卸载,其他未被

覆盖的用户只能进行本地计算。任意一个 MEC 服务区域内的用户数量 N 是均值为 $\pi r_0^2 \lambda_u$ 的泊松随机变量。对于任意一个用户,可供其进行计算卸载的服务器数量 W 是均值为 $\pi r_0^2 \lambda_s$ 的泊松随机变量。当同一用户被多个 MEC 服务器覆盖时,它将把计算任务同时卸载到多个服务器进行计算以提高计算成功概率(其中有一个服务器完成计算,即代表计算成功)。



▲图1 MEC系统模型



▲图2 大规模 MEC 网络空间分布

1.2 边缘计算卸载模型

1) 计算卸载比例

所有在 MEC 服务区域内的用户,均可决定是进行计算卸载还是进行本地计算。将网络中的选择计算卸载的用户所占比例称作用户计算卸载比例,表示为 $\rho \in (0,1)$ 。若 $\rho = 1$,则表示全部被服务区域覆盖的用户均选择计算卸载。因此,网络中选择计算卸载的用户密度表示为 $(1 - e^{-\lambda_u \pi r_0^2}) \rho \lambda_u$ 。

2) 上行信道模型

考虑用户计算卸载时的上传信道,忽略功率控制影响,将 MEC 服务器端接收到的用户信号功率表示为 $qg|Y - X|^{-\alpha}$ 。其中, q 为发射功率, g 为小尺度锐利衰落系数, α 为大尺度路径损耗系数, $|Y - X|$ 为用户 Y 与服务器 X 之间的欧氏距离。假设在任意一子信道 $m \in \{1,2,\dots,M\}$ 内,进行计算卸载的用户数服从均值为 $\frac{(1 - e^{-\lambda_u \pi r_0^2})}{M} \rho \lambda_u$ 的 PPP 分布 $\Phi_u^{(m)} = \{Y_m\} \subseteq \Phi_u$, 对于选择第 m 个子信道进行计算卸载的用户所受到的干扰信号 I 可以表示为 $I = \sum_{Y_{x,m} \in \Phi_u^{(m)} \setminus \{Y_{x,m}^*\}} qg_m |Y_{x,m}|^{-\alpha}$, 其中 g_m 服从独立恒等分布。我们对任意选取的一个典型用户进行分析,在 MEC 服务器端,该典型用户信号的信号干扰比 SIR 可以表示为 $SIR = \frac{g_0 |Y_{x,m}^*|^{-\alpha}}{\sum_{Y_m \in \Phi_u^{(m)} \setminus \{Y_{x,m}^*\}} g_m |Y_m|^{-\alpha}}$, 其中,当 SIR 不低于固定阈值 θ 时,则为卸载传输成功。

1.3 计算模型

在计算过程中,主要考虑两个约

束条件:一是时延约束 T_s ,即每个用户的计算任务需要在 T_s 秒内计算完毕;二是能量约束 ξ ,即每个用户在每个时隙用于计算的能量不得超过 ξ 。为便于分析,令 $\xi = qT_s$ 以保证每个用户均有足够的能量用于计算卸载。

1) 边缘计算

假设每个 MEC 服务器的计算能力有限,每当其接收到一个用户卸载的计算任务(包含 ℓ 比特数据),便启动一个虚拟机进行独立的边缘计算。对于进行计算卸载用户,其每个计算任务的时延包括 3 部分:卸载传输时延 T_t 、边缘计算时延 T_c 、计算结果下载时延 T_d 。由于计算结果数据量很小,因此可忽略 T_d 。为满足时延约束条件,要求 $T_t + T_c \leq T_s$ 。对于卸载传输时延 T_t ,令用户数据传输速率为 $\eta = B \cdot \log_2(1 + \theta)$,其中 B 为子信道带宽, T_t 可表示为 $T_t = \ell/\eta$,期间能量消耗为 qT_t 。对于边缘计算时延 T_c ,根据文献[13]中的计算时延模型, T_c 可表示为 $T_c = T_0(1 + d)^{i-1}$,其中, i 为虚拟机数量, $d \geq 0$ 为多个虚拟机的复用退化因子, $T_0 = \ell/\mu_{ec}$ 为单个虚拟机计算每个任务的时延(μ_{ec} 是虚拟机计算能力,单位为比特/秒)。

2) 本地计算

令 μ_{lc} (单位为比特/秒) 为用户终端的本地计算算力。为满足时延约束条件,需 $\ell/\mu_{lc} \leq T_s$,等价于 $\ell \leq T_s \mu_{lc}$,即 $T_s \mu_{lc}$ 是最大本地计算数据量。令 τ_{lc} 为本地计算每比特数据所消耗的能量,为满足能量约束条件,最大本地计算数据量为 qT_s/τ_{lc} 。综上所述,最大本地计算数据量可表示为 $\ell_{lc}^{(max)} = \min\{T_s \mu_{lc}, qT_s/\tau_{lc}\}$ 。

1.4 性能指标

1) 卸载传输成功概率

为定量刻画上传信道(即卸载传

输)的可靠性,定义卸载传输成功概率 $p_{\ell,s}$,数学表达式为 $p_{\ell,s} = \Pr(SIR \geq \theta)$ 。

2) MEC 成功概率

首先,为刻画用户的计算任务,可以在规定时间 T_s 内计算完毕的概率 p_c ,数学表达式为 $p_c = \Pr(T_t + T_c \leq T_s)$ 。考虑到每个用户可以将计算任务卸载至其附近的 W 个 MEC 服务器,当其中任意一个 MEC 服务器能够在规定时间内完成计算任务,就意味着该用户的卸载计算成功。基于此,为衡量 MEC 服务成功概率,定义 MEC 成功概率 $p_{mec}(W)$ 为 $p_{mec}(W) = 1 - (1 - p_c p_{\ell,s})^W$ 。

3) MEC 网络空间吞吐量

为刻画大规模网络中成功完成计算的用户空间密度,定义 MEC 网络空间吞吐量 $C = C_{ec} + C_{lc}$,其中 C_{ec} 表示利用边缘计算完成的吞吐量, C_{lc} 表示利用本地计算完成的吞吐量,其数学表达式分别为:

$$C_{ec} = \mathbf{E} \left[\rho \lambda_u (1 - e^{-\lambda_u \pi r_0^2}) \cdot p_{mec}(W) \right], \quad (1)$$

$$C_{lc} = \lambda_u \left[(1 - \rho) (1 - e^{-\lambda_u \pi r_0^2}) + e^{-\lambda_u \pi r_0^2} \right] \cdot I(\ell_{lc}^{(max)} \geq \ell), \quad (2)$$

其中, $I(A)$ 为指示函数,即当事件 A 发生时, $I(A)$ 为 1,否则为 0。

结合公式(1)和(2),可以得到 C 的表达式:

$$C = \lambda_u \left[(1 - \rho) (1 - e^{-\lambda_u \pi r_0^2}) + e^{-\lambda_u \pi r_0^2} \right] \cdot I(\ell_{lc}^{(max)} \geq \ell) + \mathbf{E} \left[\rho \lambda_u (1 - e^{-\lambda_u \pi r_0^2}) p_{mec}(W) \right]. \quad (3)$$

2 对 MEC 网络空间计算吞吐量的理论分析

在本节中,我们将对 MEC 成功

概率以及 MEC 网络空间计算吞吐量等关键性能指标进行分析,为后续对网络进行优化设计提供理论基础。

2.1 卸载传输成功概率分析

鉴于直接得到 $p_{\ell,s}$ 的表达式存在难度,故先推导出 $p_{\ell,s}$ 的下界来分析最差情况下的卸载传输成功概率,这同样对网络优化设计具备参考意义。首先,假设典型用户与 MEC 服务器间的距离为 r_0 ,可得到 SIR 的下界 $\text{SIR} \geq \frac{g_0 r_0^{-\alpha}}{\sum_{Y_m \in \Phi_u^{(m)} \setminus \{Y_m^*\}} g_m |Y_m|^{-\alpha}} = \text{SIR}^{(\text{low})}$, 由此可得到 $p_{\ell,s}$ 的下界:

$$p_{\ell,s} \geq \exp\left[-\frac{2\pi}{\alpha}\left(1 - e^{-\lambda_u \pi r_0^2}\right)\left(\theta r_0^2\right)^{\frac{2}{\alpha}}\left(\frac{\rho \lambda_u}{M}\right)B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right)\right] = p_{\ell,s}^{(\text{low})}, \quad (4)$$

其中 $B(x,y) = \int_0^1 \kappa^{x-1} (1-\kappa)^{y-1} d\kappa$ 。

从公式(4)中可观察到,当网络中卸载用户密度增大时(即 λ_u 或 ρ 增大),将引起更为严重的用户间干扰,致使 $p_{\ell,s}^{(\text{low})}$ 变小,扩大 MEC 服务范围 r_0 有类似效果。

2.2 MEC 成功概率分析

首先计算分析 p_c 。根据 MEC 时延约束,在 MEC 服务器中每个时隙可产生的虚拟机最大数目 $i^{(\text{max})}$ 为 $1 \leq i \leq$

$$\left\lfloor \frac{\ln\left(\frac{T_s \eta - \ell}{T_0 \eta}\right)}{\ln(1+d)} \right\rfloor + 1 = i^{(\text{max})}, \text{ 其中, } x \text{ 为下取整函数。}$$

在每个时隙中,每个 MEC 服务范围内有平均 $\rho \tilde{N} = \rho N p_{\ell,s}$ 个用户成功地将计算任务卸载到服务器端,为保证

每个任务都得到计算,需要 $\rho \tilde{N} \leq i^{(\text{max})}$,因此 p_c 可改写为:

$$p_c = \Pr(\rho \tilde{N} \leq i^{(\text{max})}). \quad (5)$$

由于 $\rho \tilde{N}$ 是均值为 $\rho \lambda_u \pi r_0^2 p_{\ell,s}$ 的泊松随机变量,故 $\Pr(\rho \tilde{N} = k) = \frac{(\rho \lambda_u \pi r_0^2 p_{\ell,s})^k}{k!} \cdot \exp(-\rho \lambda_u \pi r_0^2 p_{\ell,s})$, 因此 p_c 可表示为:

$$p_c = \sum_{k=0}^{i^{(\text{max})}} \frac{(\rho \lambda_u \pi r_0^2 p_{\ell,s})^k}{k!} \exp(-\rho \lambda_u \pi r_0^2 p_{\ell,s}) = \frac{\Gamma(i^{(\text{max})} \rho \lambda_u \pi r_0^2 p_{\ell,s})}{\Gamma(i^{(\text{max})})}, \quad (6)$$

基于公式(6), p_c 下界可表示为:

$$p_c \geq \frac{1}{\Gamma(i^{(\text{max})})} (\rho \lambda_u \pi r_0^2 p_{\ell,s}^{(\text{low})})^{i^{(\text{max})}-1} \exp(-\rho \lambda_u \pi r_0^2 p_{\ell,s}^{(\text{low})}) = p_c^{(\text{low})}, \quad (7)$$

基于公式(4)和(7),可得到 MEC 成功概率的下界:

$$p_{\text{mec}}(W) \geq 1 - \left(1 - p_c^{(\text{low})} \cdot p_{\ell,s}^{(\text{low})}\right)^W = 1 - \left[1 - \frac{1}{\Gamma(i^{(\text{max})})} \left(\rho \lambda_u \pi r_0^2 e^{-\frac{2\pi \rho \lambda_u (\theta r_0^2)^{\frac{2}{\alpha}} \beta \left(\frac{2}{\alpha} - \frac{2}{\alpha}\right)}{\alpha M} \left(1 - e^{-\lambda_u \pi r_0^2}\right)}\right)^{i^{(\text{max})}-1} \times \exp\left(\rho \lambda_u \pi r_0^2 e^{-\frac{2\pi \rho \lambda_u (\theta r_0^2)^{\frac{2}{\alpha}} \beta \left(\frac{2}{\alpha} - \frac{2}{\alpha}\right)}{\alpha M} \left(1 - e^{-\lambda_u \pi r_0^2}\right)}\right)\right]^W, \quad (8)$$

观察公式(8),假设 $p_c^{(\text{low})} \cdot p_{\ell,s}^{(\text{low})} \rightarrow 0$, 将有 $p_{\text{mec}}(W) \approx W \cdot (p_c^{(\text{low})} \cdot p_{\ell,s}^{(\text{low})})$ 。从中可见,用户的 MEC 成功概率会随着周围的服务器数量的增多而增长,符合直观预期。

根据公式(3)和公式(8),MEC 网络空间计算吞吐量的下界则可以表示为:

$$C \geq \rho \lambda_u \left(1 - e^{-\lambda_u \pi r_0^2}\right) \left(1 + e^{-\lambda_u \pi r_0^2} - e^{-\lambda_u \pi r_0^2} p_c^{(\text{low})} p_{\ell,s}^{(\text{low})}\right) + \lambda_u \left[\left(1 - \rho\right) \left(1 - e^{-\lambda_u \pi r_0^2}\right) + e^{-\lambda_u \pi r_0^2}\right] \cdot I(\ell_{lc}^{(\text{max})} \geq \ell). \quad (9)$$

讨论1:观察公式(9),若当用户具备足够的本地能量和计算能力,即 $I(\ell_{lc}^{(\text{max})} \geq \ell) = 1$,使用计算卸载将不会带来对网络计算吞吐量的增益, C 的下界将随着 r_0 和 ρ 的增加而线性减小。当用户本地能量和计算能力不足而选择计算卸载时,即 $I(\ell_{lc}^{(\text{max})} \geq \ell) = 0$,可以通过设计最优的 r_0 和 ρ 来最大化网络计算吞吐量,最优设计见第3节。

3 对 MEC 网络空间计算吞吐量的优化设计

当用户选择计算卸载时,即 $I(\ell_{lc}^{(\text{max})} \geq \ell) = 0$,通过设计最优的 MEC 服务范围 r_0 以及计算卸载比例 ρ 可以使 MEC 网络空间吞吐量 C 最大化。

3.1 优化 MEC 服务范围半径

当 $I(\ell_{lc}^{(\text{max})} \geq \ell) = 0$,根据公式(9), C 的下界可简化为:

$$C \geq \rho \lambda_u \left(1 - e^{-\lambda_u \pi r_0^2}\right) \left(1 + e^{-\lambda_u \pi r_0^2} - e^{-\lambda_u \pi r_0^2} p_c^{(\text{low})} p_{\ell,s}^{(\text{low})}\right) = C^{(\text{low})}. \quad (10)$$

观察公式(10)并考虑优化 r_0 的物理含义。一方面,当 r_0 变小时(即每个 MEC 服务范围变小),每个 MEC 服务器接收到的卸载计算任务数量会减小,由此将缩短对每个任务的计算时延从而提升 MEC 成功概率;另一方面,当 r_0 变大时,每个用户将会被更多的 MEC 服务器所覆盖并有更大概率实现 MEC,由此 MEC 成功概率也会提升。因此,我们可以设计最优

的 r_0 使 C 最大化。

由于直接根据公式(10)来优化 r_0 存在一定难度,首先考虑两种特殊情况下的优化设计,情况1是假设上行信道十分可靠时,以致卸载传输成功概率为1(即 $p_{\ell_s} = 1$),情况2是当MEC服务器计算能力很强时,以致卸载任务总能在规定时间内完成计算(即 $T_0 \rightarrow 0, p_c = 1$)。

首先考虑情况1,当 $p_{\ell_s} = 1, C^{(low)}$ 表示为:

$$C^{(low)} = \rho \lambda_u \left(1 - e^{-c_1 r_0^*}\right) \left(1 + e^{-c_1 r_0^*} - e^{-c_1 r_0^{*2} p_c^{(low)}}\right), \quad (11)$$

其中 $c_1 = \lambda_b \pi$ 。由于 $C^{(low)}$ 对 r_0 是可微的,通过优化设计 r_0 来最大化 $C^{(low)}$ 的问题可以表示为P1:

$$P1: \max_{r_0 \in \mathbb{R}^+} \rho \lambda_u \left(1 - e^{-c_1 r_0^*}\right) \left(1 + e^{-c_1 r_0^*} - e^{-c_1 r_0^{*2} p_c^{(low)}}\right). \quad (12)$$

P1为简单的凸优化问题,对P1的最优解 r_0^* 可通过求解等式(13)得到:

$$\left(r_0^* (1 - p_c^{(low)})' - 2p_c^{(low)} - 2\right) e^{c_1 r_0^{*2} (1 - p_c^{(low)})} = \left(r_0^* (1 - p_c^{(low)})' - 2p_c^{(low)}\right) e^{c_1 r_0^{*2} (2 - p_c^{(low)})} - 4, \quad (13)$$

其中, $p_c^{(low)} = c_2 (r_0^*)^{2c_1^{(max)} - 2} \exp(-c_2 r_0^*)$, $(1 - p_c^{(low)})' = 2c_2 (r_0^*)^{2c_1^{(max)} - 3} \exp(-c_2 r_0^*) (c_2 (r_0^*)^2 - i^{(max)} + 1)$, c_1 和 c_2 为常量,分别为 $c_1 = \frac{(\rho \lambda_u \pi)^{i^{(max)} - 1}}{\Gamma(i^{(max)})}$ 。求解等式(13)可以利用MATLAB等常用软件工具。

当MEC服务器密度极高时,即 $\lambda_b \rightarrow \infty$,P1可简化为P2:

$$P2: \max_{r_0 \in \mathbb{R}^+} \rho \lambda_u \left(1 - e^{-c_1 r_0^* p_c^{(low)}}\right). \quad (14)$$

对P2的最优解 r_0^* 可表示为如下

$$\text{闭式解 } r_0^* = \left(\frac{i^{(max)}}{\rho \lambda_u \pi}\right)^{\frac{1}{2}}.$$

讨论2:观察 r_0^* 的闭式解,发现 r_0^* 会随着最大可产生的虚拟机数量 $i^{(max)}$ 的增加而变大,这是因为 $i^{(max)}$ 越大意味着MEC服务器计算能力越强,从而能提供更大范围的服务。同时, r_0^* 会随着卸载用户密度 $\rho \lambda_u$ 的增加而变小,是因为用户卸载的计算任务越多,将会给MEC服务器带来更大计算压力,故需要缩小MEC服务范围来减小计算压力,用以保证一定的MEC成功概率。

接下来,考虑情况2。当 $p_c = 1, C^{(low)}$ 可表示为:

$$C^{(low)} = \rho \lambda_u \left(1 - e^{-c_1 r_0^*}\right) \left(1 + e^{-c_1 r_0^*} - e^{-c_1 r_0^{*2} p_c^{(low)}}\right). \quad (15)$$

优化设计 r_0 的问题可表示为P3:

$$P3: \max_{r_0 \in \mathbb{R}^+} \rho \lambda_u \left(1 - e^{-c_1 r_0^*}\right) \left(1 + e^{-c_1 r_0^*} - e^{-c_1 r_0^{*2} p_c^{(low)}}\right). \quad (16)$$

对P3的最优解 r_0^* 可通过求解等式(17)得到:

$$\left(2p_{\ell_s}^{(low)} + r_0^* (p_{\ell_s}^{(low)})' + 2\right) e^{c_1 (r_0^*)^2 (1 - p_{\ell_s}^{(low)})} = \left(2p_{\ell_s}^{(low)} + r_0^* (p_{\ell_s}^{(low)})'\right) e^{c_1 (r_0^*)^2 (2 - p_{\ell_s}^{(low)})} - 4, \quad (17)$$

其中, $p_{\ell_s}^{(low)} = \exp\left(-c_3 (r_0^*)^2 \left(1 - e^{-c_1 (r_0^*)^2}\right)\right)$, $(p_{\ell_s}^{(low)})' = 2c_3 \left(r_0^* \left(e^{-c_1 (r_0^*)^2} - 1\right) - c_1 (r_0^*)^3 e^{-c_1 (r_0^*)^2}\right) \exp\left(-c_3 (r_0^*)^2 \left(1 - e^{-c_1 (r_0^*)^2}\right)\right)$, 其中, $c_1 = \lambda_b \pi$, $c_3 = \frac{2\pi\rho\lambda_b\theta^\alpha}{\alpha M} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right)$ 。

当MEC服务器密度极高时(即

$\lambda_b \rightarrow \infty$),P3可简化为P4:

$$P4: \max_{r_0 \in \mathbb{R}^+} \lambda_u \left(1 - e^{-c_1 r_0^* p_c^{(low)}}\right). \quad (18)$$

对P4的最优解 r_0^* 可表示为如下

$$\text{闭式解 } r_0^* = \left(\frac{2\pi\rho\lambda_b\theta^\alpha}{\alpha M} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right)\right)^{-\frac{1}{2}}.$$

3.2 优化计算卸载比例

首先考虑通过优化 ρ 来最大化MEC网络空间吞吐量 C 的物理含义。太大或太小的 ρ 都会降低 C ,这是因为:太大的 ρ 将会引起更严重的用户间干扰和卸载计算任务数量,导致MEC成功概率降低,从而降低 C ;太小的 ρ 会直接降低网络中卸载用户密度,导致 C 的减小。因此,可以通过优化 ρ 来最大化 C 。

类似3.1节的步骤,首先考虑当上行信道十分可靠时(即 $p_{\ell_s} = 1$), $C^{(low)}$ 可表示为:

$$C^{(low)} = \rho \lambda_u \left(1 - e^{-c_4}\right) \left(1 + e^{-c_4} - e^{-c_4 p_c^{(low)}}\right), \quad (19)$$

其中 $c_4 = \lambda_b \pi r_0^2$ 。由于 $C^{(low)}$ 对 ρ 是可微的,通过优化设计 ρ 来最大化 $C^{(low)}$ 的问题可以表示为P5:

$$P5: \max_{\rho \in [0,1]} \rho \left(1 + e^{-c_4} - e^{-c_4 p_c^{(low)}}\right), \quad (20)$$

P5为凸优化问题,对P5的最优解 ρ^* 可通过求解等式(21)得到:

$$\rho^* c_4 (p_c^{(low)})' e^{c_4 (1 - p_c^{(low)})} = e^{c_4 (1 - p_c^{(low)})} - 2, \quad (21)$$

其中, $p_c^{(low)} = c_6 (\rho^*)^{i^{(max)} - 1} e^{-c_5 \rho^*}$, $(p_c^{(low)})' = c_6 (\rho^*)^{i^{(max)} - 2} e^{-c_5 \rho^*} (i^{(max)} - c_5 \rho^* - 1)$, $c_4 = \lambda_b \pi r_0^2$, $c_5 = \lambda_u \pi r_0^2$, 以及 $c_6 = \frac{c_5^{i^{(max)} - 1}}{\Gamma(i^{(max)})}$ 。

为了得到物理含义更明显的结果,利用不等式 $1 - e^{-x} \leq x$,对公式(19)进行简化,可得到 $C^{(low)}$ 的近似结

果, $C^{(low)} \approx \rho \lambda_u (1 - e^{-c_4}) (c_4 p_c^{(low)} + e^{-c_4})$ 。
由此,当 MEC 服务器密度极高时(即 $\lambda_b \rightarrow \infty$),最优化问题可设计为 P6:

$$P6: \max_{\rho \in [0,1]} c_4 \rho p_c^{(low)} \quad (22)$$

对 P6 的最优解 ρ^* 为如下闭式解 $\rho^* = \frac{i^{(max)}}{\lambda_u \pi r_0^2}$ 。

讨论3:观察 ρ^* 闭式解,当 $i^{(max)}$ 增加时,表示 MEC 服务器计算能力强,便可提高 ρ^* 以加大卸载用户数目,来提高网络计算吞吐量;当用户密度 λ_u 增加时,则应减小 ρ^* 以减轻 MEC 服务器计算压力,从而保证一定的 MEC 成功概率。

接下来,考虑当 MEC 服务器计算能力很强(即 $p_c = 1$), $C^{(low)}$ 可表示为:

$$C^{(low)} = \rho \lambda_u (1 - e^{-c_4}) (1 + e^{-c_4} - e^{-c_4 \rho_{\ell,s}^{(low)}}) \quad (23)$$

基于公式(23),对于 ρ^* 的最优化问题可设计为 P7:

$$P7: \max_{\rho \in [0,1]} \rho (1 + e^{-c_4} - e^{-c_4 \rho_{\ell,s}^{(low)}}) \quad (24)$$

对 P7 的最优解 ρ^* 可通过求解等

式(25)得到:

$$\rho^* c_4 (p_{\ell,s}^{(low)})' e^{c_4(1-p_{\ell,s}^{(low)})} = e^{c_4(1-p_{\ell,s}^{(low)})} - 2, \quad (25)$$

其中, $p_{\ell,s}^{(low)} = e^{-c_7 \rho^*}$, $(p_{\ell,s}^{(low)})' = -c_7 e^{-c_7 \rho^*}$,

$$c_7 = \frac{2c_5(1 - e^{-c_4})\theta^\alpha}{\alpha M} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right)$$

当 MEC 服务器密度极高时(即 $\lambda_b \rightarrow \infty$), P7 可简化为如下优化问题:

$$\max_{\rho \in [0,1]} c_4 \rho p_{\ell,s}^{(low)}, \quad (26)$$

其最优解 ρ^* 可表示为如下闭式

$$\rho^* = \frac{\alpha M}{2\lambda_u \pi r_0^2 \theta^\alpha B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right)}$$

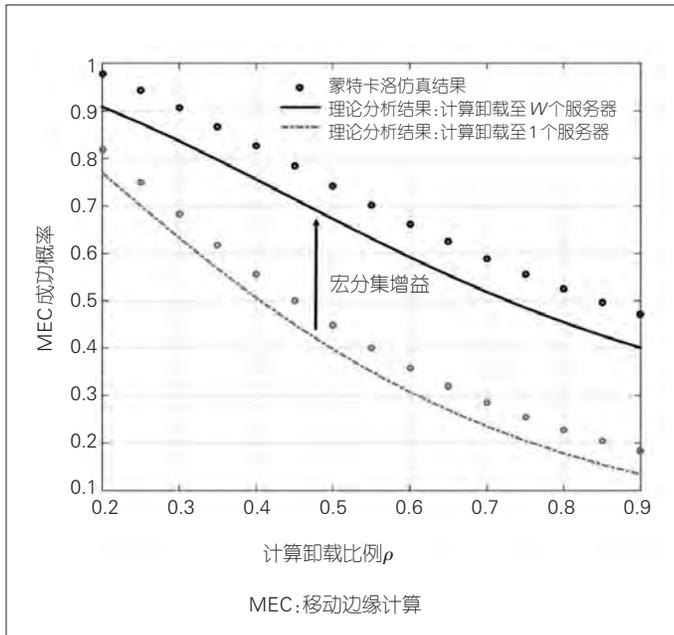
4 仿真结果

在本节中,我们利用 MATLAB 仿真对上文中得到的理论结果加以验证,主要仿真参数设置如下: $\lambda_b = 0.01/\text{m}^2$, $\lambda_u = 0.1/\text{m}^2$, $r_0 = 8 \text{ m}$, $\rho = 0.7$, $\theta = 10 \text{ dB}$, $B = 3 \text{ kHz}$, $\alpha = 3$, $T_s = 100 \text{ ms}$, $\ell = 10^3 \text{ bits}$, $T_0 = 1 \text{ ms}$, $d = 0.3$ 。其中,

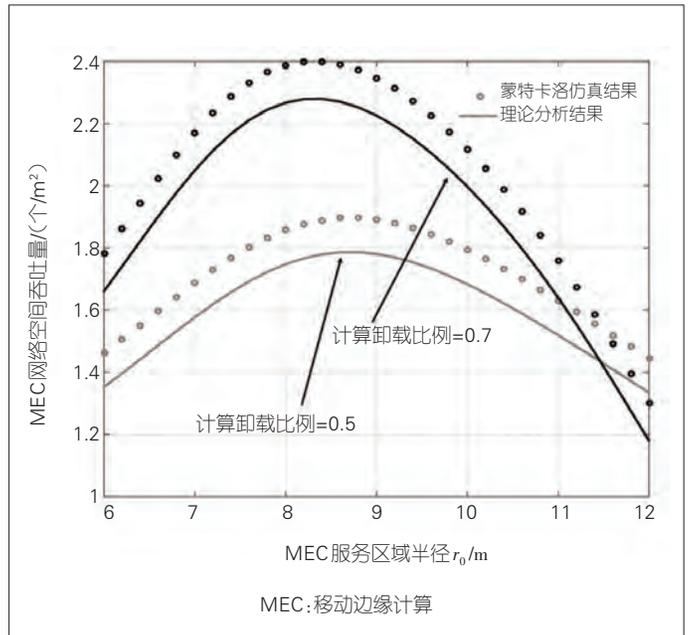
蒙特卡洛仿真结果由圆圈表示,理论分析结果由实/虚线表示。限于篇幅,这里只展示最重要的3个仿真结果。

图3展示了典型用户的 MEC 成功概率,其中,虚线表示用户将计算任务只卸载到任意一个服务器时 MEC 成功概率,实线则表示用户将计算任务卸载到附近的 W 个服务器时的 MEC 成功概率,即公式(8)。首先,文中得到的理论结果(下界)与仿真结果之间的差值较小,这证明理论结果比较准确;其次,可以观察到,相比选择一个 MEC 服务器进行计算卸载,当用户选择向 W 个服务器同时进行计算卸载时的 MEC 成功概率有明显提升,这得益于宏分集增益。

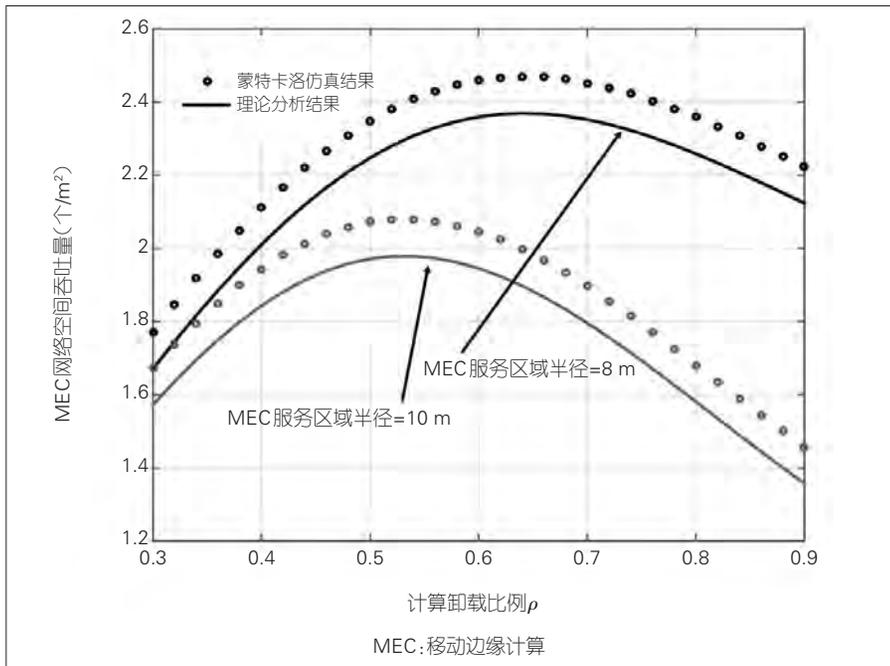
图4展示了通过优化 r_0 来最大化 MEC 网络吞吐量 C 。首先,文中我们所求得的 $C^{(low)}$ (下界)相比 C 仅有少量差值,这表示理论下界较为准确;其次,当给定 ρ 的值, C 及 $C^{(low)}$ 是变量 r_0 的凹函数,因此可以通过设计最优的 r_0^* 来最大化 C ,如当 ρ 在 $0.5 \sim 0.7$ 之间时, r_0^* 在 $8 \sim 9 \text{ m}$ 之间。另外,当提高 ρ 的值, C 的最大值会随之变大,这是因



▲图3 MEC成功概率与计算卸载比例关系图



▲图4 MEC网络空间吞吐量与MEC服务区域半径关系图



▲图5 MEC成功概率与计算卸载比例关系图

为增大 ρ 意味着更多的用户选择计算卸载,从而有效增大网络吞吐量。

图5展示了通过优化 ρ 来最大化MEC网络吞吐量 C 。给定 r_0, C 及 $C^{(low)}$ 同样是变量 ρ 的凹函数。当 r_0 在8~10m之间时,最优值 ρ^* 在0.5~0.7间。另外,当 r_0 增大时,MEC服务范围将扩大,更多用户的计算任务可卸载至服务器;因此,网络吞吐量也会增加。

5 结束语

本文中,我们首次定义了大规模MEC网络中的空间计算吞吐量这一性能指标,并通过优化设计MEC服务范围半径 r_0 以及用户计算卸载比例 ρ 这两个指标,实现MEC网络空间吞吐量的最大化。所提供的理论分析与优化结果将为部署大规模MEC网络提供了极为重要的设计参考。

参考文献

[1] MACH P, BECVAR Z. Mobile edge computing: a survey on architecture and computation offloading [J]. IEEE communications surveys and tutorials, 2017, 19(3): 1628–1656. DOI: 10.1109/comst.2017.2682318

[2] WU D, WANG F, CAO X, et al. Joint communication and computation optimization for wireless powered mobile edge computing with D2D offloading [J]. Journal of communications and information networks, 2019, 4(4): 72–86

[3] 马洪源. 面向5G的边缘计算及部署思考 [J]. 中兴通讯技术, 2019, 25(3): 77–81. DOI: 10.12142/ZTETJ.201903011

[4] 丁春涛, 曹建农, 杨磊, 等. 边缘计算综述:应用、现状及挑战 [J]. 中兴通讯技术, 2019, 25(3): 2–7. DOI: 10.12142/ZTETJ.201903001

[5] YANG X, HUA S, SHI Y, et al. Sparse optimization for green edge AI inference [J]. Journal of communications and information networks, 2020, 5(1): 1–15

[6] QIN M, CHEN L, ZHAO N, et al. Power-constrained edge computing with maximum processing capacity for IoT networks [J]. IEEE Internet of Things journal, 2019, 6(3): 4330–4343. DOI:10.1109/jiot.2018.2875218

[7] QIN M, CHEN L, ZHAO N, et al. Computing and relaying: utilizing mobile edge computing for P2P communications [J]. IEEE transactions on vehicular technology, 2020, 69(2): 1582–1594. DOI:10.1109/tvt.2019.2956996

[8] YOU C, HUANG K, CHAE H, et al. Energy-efficient resource allocation for mobile-edge computation offloading [J]. IEEE transactions on wireless communications, 2017, 16(3): 1397–1411. DOI:10.1109/twc.2016.2633522

[9] MAO Y, ZHANG J, LETAIEF K. Dynamic computation offloading for mobile-edge computing with energy harvesting devices [J]. IEEE communications surveys and tutorials, 2016, 14(12): 3590–3605. DOI:10.1109/jsac.2016.2611964

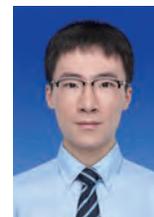
[10] ZHAO J, LI Q, GONG Y, et al. Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks [J]. IEEE transactions on vehicular technology, 2019, 68(8): 7944–7956. DOI: 10.1109/tvt.2019.2917890

[11] KO S, HAN K, HUANG K. Wireless networks for mobile edge computing: spatial modeling and latency analysis [J]. IEEE transactions on wireless communications, 2018, 17(8): 5225–5240. DOI:10.1109/twc.2018.2840120

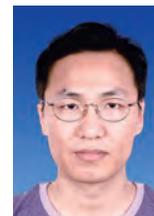
[12] HAENGGI M. Stochastic geometry for wireless networks [M]. Cambridge: Cambridge University Press, 2009. DOI: 10.1017/cbo9781139043816

[13] BRUNEO D. A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems [J]. IEEE transactions on parallel and distributed system, 2014, 25(3): 560–569. DOI:10.1109/tpds.2013.67

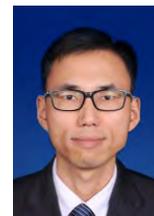
作者简介



韩凯峰, 中国信息通信研究院政策与经济研究所工程师, 工学博士; 主要研究领域为无线通信(5G/6G)、边缘计算、车联网、人工智能等前沿技术及战略; 已发表论文15篇。



胡昌军, 中国信息通信研究院政策与经济研究所战略研究部副主任、高级工程师, 工程硕士, 工信部《新一代人工智能产业创新重点任务揭榜工作方案》起草团队的主要参与者; 主要研究领域为人工智能、5G等ICT前沿领域技术及战略; 曾先后参加工信部电子发展基金等多项课题研究。



刘铁志, 中国信息通信研究院政策与经济研究所战略研究部主任、高级工程师, 经济学博士, 曾担任《国务院关于加快构建大众创业万众创新支撑平台的指导意见》、《互联网+人工智能三年行动计划(2018–2020年)》《关于促进人工智能和实体经济深度融合的指导意见》等起草专家团队的主要负责人; 主要研究领域为信息通信产业政策和国际经济。



基于神经网络计算的 无线容量高实时预测

High Real-Time Capacity Prediction Based on Neural Network Evaluation

赖昱辰 / LAI Yuchen¹, 钟祎 / ZHONG Yi¹, 王建峰 / WANG Jianfeng²

(1. 华中科技大学, 中国 武汉 430074;

2. 微软公司, 美国 雷德蒙德 98052)

(1. Huazhong University of Science and Technology, Wuhan 430074, China;

2. Microsoft Corporation, Redmond, 98052, USA)

摘要: 提出了一种基于卷积神经网络 (CNN) 计算的无线网络容量高实时预测方法。针对不同的网络部署环境, 考虑路径损耗、信道衰落、墙损等因素, 分别建立无线网络模型以获取数据集。将无线网络中接入点的部署方案视为二维矩阵像素图, 并作为神经网络的输入, 将无线网络容量标记为标签。使用 CNN 处理矩阵并输出数值, 与标签值对比进行权重优化, 再仿真验证 CNN 的不同架构和参数的影响。CNN 可以更智能和高效地进行无线网络性能的评估与优化, 实现大规模物联网 (IoT) 网络的部署和监管, 具有高准确性和鲁棒性。

关键词: 卷积神经网络; 容量预测; 网络部署; 干扰

Abstract: Based on the convolution neural network (CNN) evaluation, a high real-time prediction method for wireless network capacity is proposed. According to diversified aspects such as path loss, channel fading and wall loss, wireless network models in different deployment environments are established to obtain data sets. Then the deployment patterns of access points are regarded as 2-dimensional matrix pixel maps, which are the inputs of the neural network, and the values of the wireless capacity are marked as labels. CNN is used to handle matrices, output numeric, compare with the label value for weight optimization, and verify the performance of CNN models with different architectures and parameters through simulation. CNN can enable more intelligent and efficient wireless network performance evaluation and optimization, realize the deployment and regulation of massive Internet of Things (IoT) networks, and prove high accuracy and robustness.

Keywords: convolutional neural network; capacity prediction; network deployment; interference

DOI: 10.12142/ZTETJ.202004004

网络出版地址: <https://kns.cnki.net/KCMS/detail/34.1228.TN.20200713.1407.004.html>

网络出版日期: 2020-07-13

收稿日期: 2020-05-29

随着无线数据需求的急剧增加, 用户与无线接入点之间的距离大大减少, 导致无线网络向超密集架构发展。由于密集的部署和多样化的传播环境, 网络管理和调节也变得极为复杂。在室外环境中, 需要额外

部署小型基站, 减轻宏基站的无线流量负担, 提高覆盖范围的连续性和信号强度。在室内传播环境如购物中心和办公楼内, 由于承重墙等障碍物, 加上个体用户密集部署的无线局域网 (WLAN) 的影响, 信号传播可能比室外更为复杂。此外, 由于覆盖率、速率、等待时间的不同, 不同无线应用有着多样化的服务质量 (QoS) 要求。

因此, 应当正确规划接入点的部署, 以保证无线网络的性能。

为规划无线网络的部署, 首先应评估给定部署方案的容量。传统方法中, 网络的容量通过系统级仿真进行估算, 需要计算每个位置的信噪比 (SINR) 值, 再计算每个位置的数据率; 然而, 模拟具有复杂传播环境的大规模无线网络相当复杂, 且仅适用于特

基金项目: 国家自然科学基金 (61701183)、中央高校基本科研业务费专项资金 (2018KFYYXJJ139)

定的部署方案,而部署方案千变万化,所以对网络容量的评估应该更加智能和高效。

本文中,我们提出了使用卷积神经网络(CNN)模型预测无线网络的容量的方法。图1中给出了容量预测模型的框架,由用于无线网络仿真的系统模型和用于数据预测的CNN组成。首先,在方形区域和有墙环境中使用无线网络模型模拟信号传输,将环境信息与接入点信息等多维数据应用于网络容量估计,使用二维矩阵描述接入点部署位置,并将无线网络容量标记为标签,得到CNN数据集。其次,将数据集按一定比例分为训练集和测试集两部分,建立CNN模型并将训练集导入,经过特征提取与分类决策后输出网络容量值,再利用反向传播算法动态调整神经网络参数的权重,优化数据预测模型。

1 无线网络系统建模

在本节中,我们介绍两种不同的无线网络部署环境(方形无障碍区域与有墙环境)并建立了信号传输模型。

在两个环境中各自生成10 000个接入点部署图并计算其网络容量,作为卷积神经网络的数据集。

我们将区域均匀划分为个 $M \times M$ 网格,随机分布 N 个发射功率相同的接入点。区域中,有 M^2 个用户均匀分布在平面内。在方形区域内,每个用户连接与其距离最近的接入点。在有墙环境中,用户连接到穿墙后能提供最大信号功率的接入点,并将其他接入点视作干扰。对每个接入点发射的信号,需要考虑的关键因素有:

1) 路径损耗。该损耗是由发射功率的辐射扩散和无线信道的传播特性引起的。电波信号随距离增长而衰减,根据标准信号传播模型,设定路径损耗系数大于2。

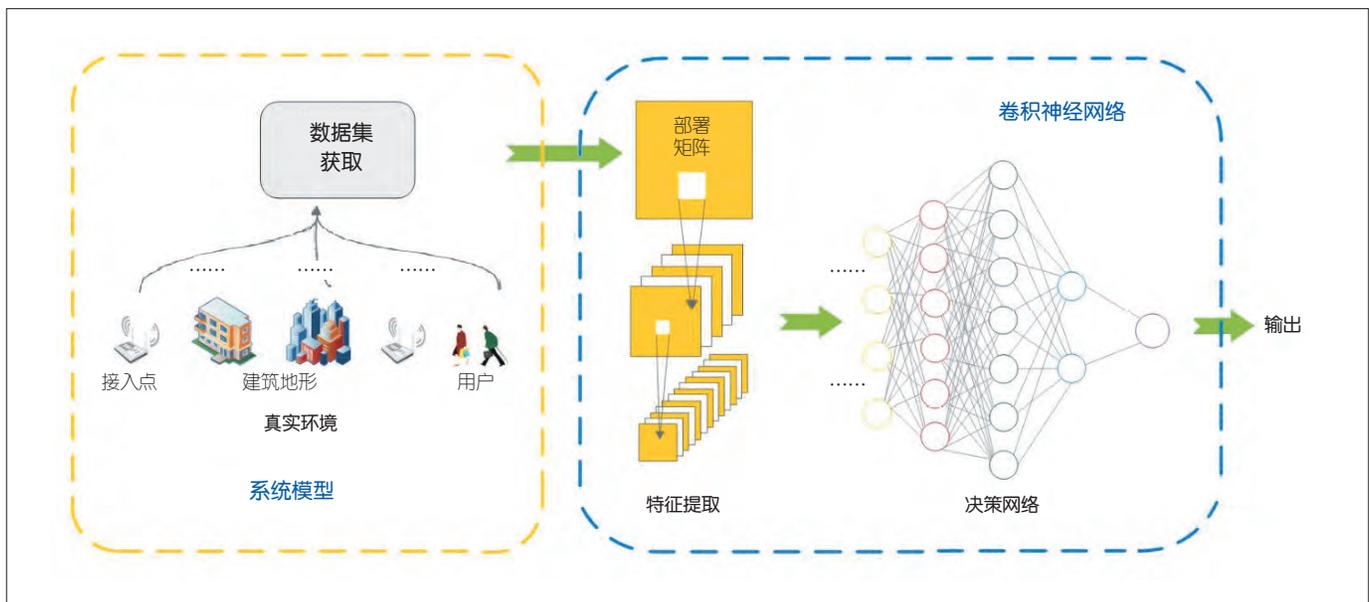
2) 瑞利衰落。由于信号的反射、衍射、折射和散射引起无线信道的多径传播,使得接收到的信号幅度遵循瑞利分布,功率衰减系数遵循均值为1的指数分布。

3) 墙损。对有墙环境,还需要考虑穿墙带来的信号损耗。信号通过一面墙时的衰减值为10~15 dB,故

设定墙损系数为0.314。此外,采用跨立实验方法来计算在用户和接入点之间的无线通信链路上的墙的数量。

与接收信号功率相比,通信环境中的高斯噪声的干扰可以忽略不计,利用香农公式可以计算每个用户的网络容量值。在方形区域内需考虑到边缘效应的影响,在4条边界各忽略15%的区域,累计中心区域的用户的数据率,取平均得到全局网络容量值。对有墙环境,则累加平面图内所有室内区域的数据率并取平均值。

图2给出了一个室内设计方案示例图。原点设置在左下位置,将整个区域按比例缩放到平面直角坐标系中,同时将墙壁模拟为线性线段,建模得10 000个样本。图3中是获得最大和最小网络容量值的接入点的部署图和相应的全局数据率分布。其中,高网络容量值区域显示为黄色,低容量区域显示为蓝色。可以看出,数据率随着接入点周围半径的增加而减小,并在靠近接入点的区域达到峰值,但当接入点分布过于密集时会互相干扰,使得网络容量值较小。



▲图1 基于卷积神经网络的网络容量预测架构

2 基于 CNN 的容量预测

本节中，我们将分析不同的卷积神经网络架构，以探讨网络容量预测问题在不同场景下的适用性。将接入点的位置矩阵（即像素值为 $M \times M$ 的二维图像）作为 CNN 的输入，接入点位置的像素值标准化为 1，其他区域像素值设置为 0；将网络容量从大到小均分为 40 类，作为 CNN 的标签。

表 1 给出了使用的卷积神经网络的结构，Conv(x, y, z, s) 表示卷积层，其输入通道数为 x 、输出通道数为 y 、步长为 s ，卷积核的大小为 $(z \times z)$ 。MaxPool(z, s) 表示最大池化层，其卷积核大小为 $(z \times z)$ ，步长为 s 。Fc(x, y) 表示具有输入节点数为 x 与输出节点数为 y 的全连接层。

2.1 特征提取

在卷积层中，通过将卷积核连接到输入层相邻区域中的多个神经元，自动完成输入数据集的特征提取。每个卷积层都会生成一个新的特征图，其维数等于卷积核的数量，其尺寸取决于卷积核的大小和步长。通过连续卷积，特征图维数增大而尺寸减小。

卷积层的输出特征图会被传输到最大池化层，以进行特征选择和信息过滤。在最大值滤波的区域中，下采样函数提取所有连接神经元的最大值。池化层用于压缩特征图并减小输出的空间大小以简化计算，也可提取主要特征以提高网络的鲁棒性。池化层中的计算方法与卷积层中相同，滤波器的参数不会经反向传播过程被修改。

2.2 分类预测

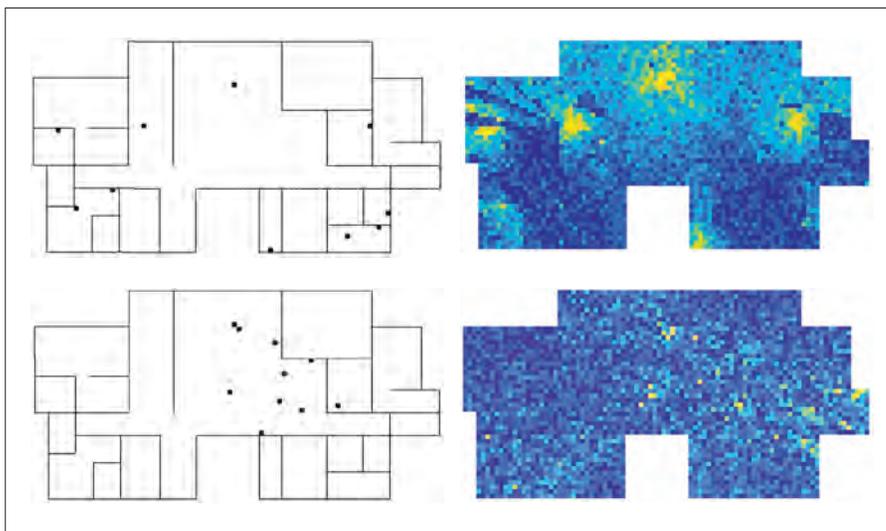
全连接层位于所有神经网络的末端，连接所有输入特征并将分散特征映射到标记的样本空间中，可用于减少特征位置对结果的影响。第一，将从最后一个卷积层获得的高维数据表

示为大小为 3、维数为 64 的特征图，并扩展成 576 个单独特征作为全连接层的输入；第二，将线性加权和应用于隐藏层，将每个标签的输出概

率发送到分类器，并在下一次训练中通过反向传播算法更新隐藏层的参数权重；第三，分类器将概率最高的标签作为最终输出。分类的数量越多，



▲图 2 有墙环境示例



▲图 3 获得最高和最低容量值的接入点的部署位置及数据率分布图

▼表 1 应用于无线网络容量预测的 CNN 的详细结构与参数信息

模型	结构
CNN-1	Conv(1,10,5,1)-Maxpool(3,2)-Conv(10,32,5,2)-Maxpool(3,2)-Conv(32,64,3,1)-Maxpool(3,2)-FC(576,240)-FC(240,40)
CNN-2	Conv(1,10,3,1)-Maxpool(2,2)-Conv(10,20,3,1)-Maxpool(3,2)-Conv(20,40,5,1)-Maxpool(3,2)-Conv(40,64,3,1)-Maxpool(3,2)-FC(576,240)-FC(240,40)
CNN-2+ BatchNorm	Conv(3,10,5,1,0)-Maxpool(3,2)-BatchNorm(10)-Conv(10,32,5,2,0)-Maxpool(3,2)-BatchNorm(32)-Conv(32,64,5,1,1)-Maxpool(3,2)-BatchNorm(64)-FC(576,240)-FC(240,40)
CNN-2+ Dropout	Conv(3,10,5,1,0)-Maxpool(3,2)-Conv(10,32,5,2,0)-Maxpool(3,2)-Conv(32,64,5,1,1)-Maxpool(3,2)-Dropout()-FC(576,240)-Dropout()-FC(240,40)

CNN: 卷积神经网络

两个相邻网络容量标签的值差就越小，即预测的网络容量的精度越高。我们设置了两个全连接层，并添加了一些非线性方法来提高数据集的训练效率。

2.3 权重更新与模型优化

卷积层与池化层具有较少的参数和较大的计算量，而全连接层则相反；因此，在加速优化过程时着重于调整卷积层的参数和结构，在实现参数优化和权重裁剪时着重于全连接层。

经过仿真测试，使用线性整流函数（ReLU）作为激活函数以解决过拟合和梯度消失的问题，同时减少计算量。使用交叉熵损失函数作为评估神经网络性能的指标，用于比较预测容量值与实际输出之间的差异。在反向传播过程中计算完所有参数的梯度后，使用基于随机梯度下降（SGD）算法的AdaGrad优化算法对网络的权重和参数进行更新，从而获得最优的权重参数。

3 仿真验证与分析

在本节中使用Pytorch框架建立所有的CNN模型，主要研究以下CNN结构和参数对训练效率与准确度的影响：分类数与数据集数量、优化算法学习率、批归一化层与Dropout层、CNN的深度。

将训练集和测试集的比例设置为6:1，为确保实验结论的普适性，每个测试重复3次以上并取结果平均值。一个时期（epoch）意味着训练集中的所有样本训练一次，且测试集的所有数据被评估一次。由于过拟合现象的产生，可以使用早停法，即提前终止训练过程以获得更高准确性。为了减少大规模样本的计算时间和梯度值差异，每32个训练集样本被划分成一个小批次，随机打乱批次顺序并分批进行训练。在以上前提条件下进行测试，观察到在方形无障碍区域内，测试集

准确率最高为96.01%，有墙环境中准确率最高为87.84%。在简单环境中应用的CNN也可以从复杂场景中提取隐藏特征，但模型训练时间更长，精度更低。

3.1 分类数与数据集数量

当方形区域的10000个数据输入到CNN-4时，准确率仅达到79.93%。增加训练集的数量可以提高CNN模型的拟合能力，当训练集的数量逐渐增加到40000个时，可基本满足准确率要求。将网络容量预测视为一个分类问题，当预测结果与真实值的误差不超过2级时可视为结果正确，也可以通过减小分类类别数提高准确率；但随着输出等级的逐渐减小，测试集的数据精度降低，经测试后选择40级输出以平衡二者性能。

3.2 神经网络深度

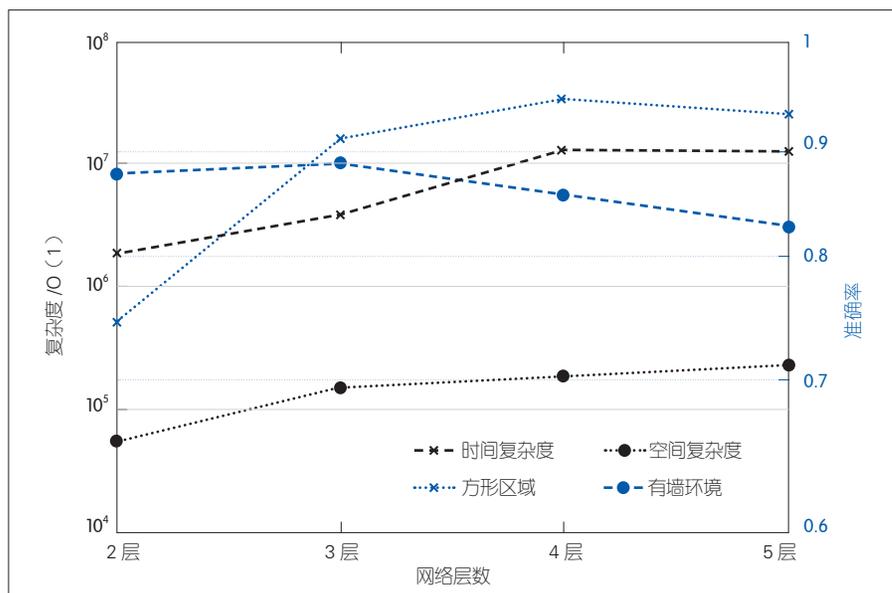
低时间复杂度意味着程序执行的语句较少，运行每个epoch的时间较短。低空间复杂度则意味着临时占用的参数和存储空间数量较少。一般情况下，模型的时间和空间复杂度随神

神经网络层数的增加而增加。由表1可知，CNN-2的模型结构与CNN-1非常相似，只是使用两层 3×3 卷积替换了CNN-1第3层卷积层的 5×5 卷积核，这使得时间复杂度反而下降。

如图4所示，在方形区域中，由于2层网络无法充分提取特征图的特征，甚至无法学习基本特征，2层CNN的准确率只能达到74.77%；但随CNN深度的增加，准确率呈上升趋势并在使用4层网络时达到最大值94.12%。在层数增加的同时，权重的线性相乘容易导致梯度爆炸或消失，且抽象能力过强时会阻止网络提取有用的功能，这将使得5层网络的准确率下降。在有墙环境中，准确率首先从2层的87.06%提高到3层的87.84%，接着持续降低。所以，在方形区域中使用4层神经网络CNN-2，在有墙环境中使用3层网络CNN-3。

3.3 优化算法学习率（LR）

图5中测试了LR介于0.001~0.1之间的Adam算法和可自动调整学习率的AdaGrad算法的性能。LR较大时，收敛速度很快，但容易出现梯度爆炸，



▲图4 2—5层卷积神经网络的时间和空间复杂度及其在方形区域与有墙环境内的准确率

使得权重更新失败,导致模型不收敛。可以看到,在 LR 为 0.1 时,模型不收敛。当 LR 为 0.01 时,准确率先上升,但梯度爆炸使得模型训练失败。LR 较小时,梯度下降慢,收敛时间较长,也可能导致过拟合问题。将 LR 减小到 0.001,模型可正常运行并表现出良好的性能。0.005 的 LR 也被测试,虽然缓慢收敛,但性能不如 0.001 的 LR。由于自适应学习率算法被越来越

多地应用,我们选择使用 AdaGrad 算法,并发现其在训练效率和准确率上显示出了更好的性能,并最终将其应用于本文的 CNN 模型内。

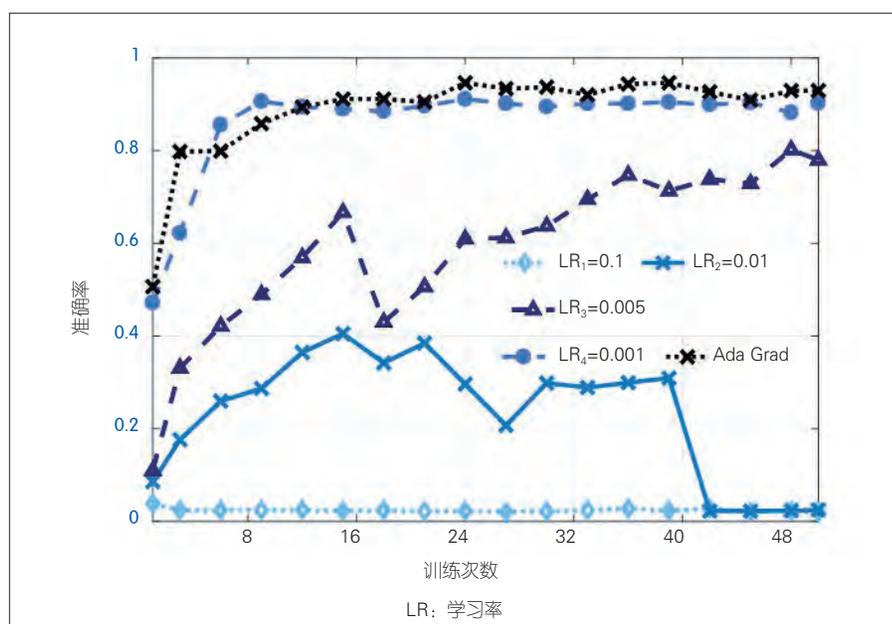
3.4 Dropout (DP) 层和批归一化 (BN) 层

BN 层将重新缩放所获均值和与方差,将每批训练数据标准化,再使用新学习的均值 0 和单位方差优化网

络梯度,使数据分布更符合训练过程中的实际情况,以确保模型的非线性。在前向传播过程中,DP 层使隐藏层的某些节点停止工作,确保该模型不会太依赖于局部特征。两者都可提高网络泛化能力,优化过拟合问题。

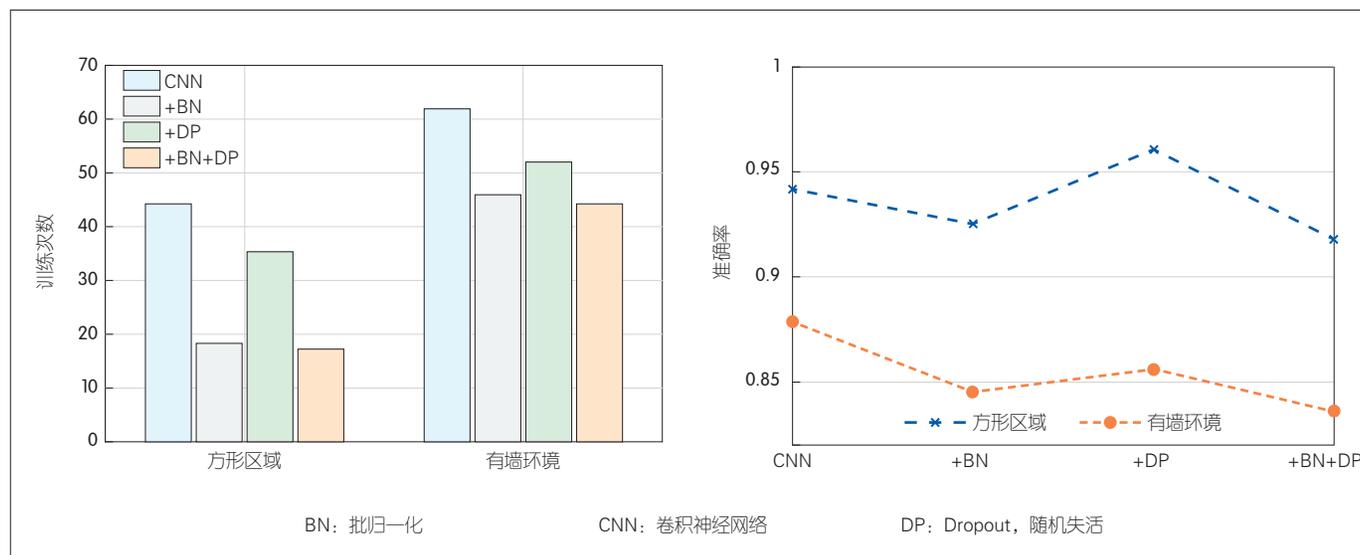
如图 6 所示,网络内不加 BN 或 DP 层时,训练次数和单次训练时长最大,但在有墙环境内的准确率最高,为 87.84 %。在每个最大池化层后添加 BN 层后,训练速度大大提高,可在第 18 和第 46 个 epoch 完成模型训练;然而,由于训练集是接入点分布的二维图像,BN 在训练过程中丢失部分特征图值,从而导致模型拟合度下降,使得准确率降低。带有 DP 层的 CNN 模型的训练速度得到了提高,同时在方形区域中的精度达到最大值 96.01 %。BN 层和 DP 层同时添加时,训练速度最高,但精度最低。

在程序中使用统一计算设备架构 (CUDA) 加速,可大大缩短训练时间,这对二者在训练效率上的影响差异可以忽略。因此在方形区域中,在 CNN-2 全连接层后加 DP 层,在有墙环境中可直接使用 CNN-1 模型。当使



▲图 5 使用 LR 为 0.001-0.1 之间的 Adam 算法和 AdaGrad 算法的准确率

下转第 22 页 ➡



▲图 6 在 DP 层和 BN 层影响下的训练次数和准确率

基于空中计算的 无线群智感知

Over-the-Air Computation Based Wireless Crowd Sensing

李晓阳/LI Xiaoyang, 贡毅/GONG Yi

(南方科技大学, 中国 深圳 518055)
(Southern University of Science and Technology, Shenzhen 518055, China)



摘要:为实现服务器对海量传感数据的快速收集,提出了一种基于空中计算的快速传感数据汇聚方案。该方案通过无线功率传输激励用户参与群智感知,为数据感知和空中计算供能。通过对无线功率分配、感知数据量,以及空中计算数据汇聚时间进行联合优化,实现服务器数据开发效益的最大化。该方案的性能通过仿真进行了验证,并与传统的无线群智感知设计进行了比较。

关键词:无线群智感知;空中计算;无线功率传输

Abstract:To achieve fast aggregation of massive sensing data at the server, an over-the-air computation (AirComp) based fast data aggregation design is proposed. In this design, wireless power transfer serves as the incentivizing mechanism for users to take part in wireless crowd sensing (WCS), as well as the energy source for data sensing and AirComp. The wireless power transfer allocation, sensing data size, and AirComp-based data aggregation time are jointly optimized to maximize the data exploitation reward of server. The simulation demonstrates the performance of the proposed design compared with the traditional WCS.

Keywords: wireless crowd sensing; over-the-air computation; wireless power transfer

DOI: 10.12142/ZTETJ.202004005
网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20200709.1708.006.html>

网络出版日期: 2020-07-09
收稿日期: 2020-05-30

物联网的迅速发展推动了数以亿计的无线传感设备部署,这些设备被用来采集各项应用所需要的数据(如温度、湿度、污染程度、车流量等)。然而,传统的无线传感器网络覆盖范围和可扩展性有限,并且存在高昂的维护成本^[1]。近年来,无线群智感知利用移动用户可穿戴设备中的传感模块,为数据采集提供了一

种新的解决方案^[2-4]。一系列奖励机制被设计用于激励用户参与群智感知,包括金钱、服务质量、用户体验等^[5]。尽管这些奖励机制起到了一定的效果,但对于无源的无线传感设备而言,更重要的一点是要有足够的电量来执行数据感知任务。为了解决这一问题,无线功率传输被设计为一种新型的奖励机制,在激励设备参与群智感知的同时能够为设备供电^[6-7]。无线功率传输最早被用于点对点的功率传输,目前已经被业界广泛应用,为各种通信系统提供了

能量^[8-10]。

然而,由于传感设备的计算能力有限,难以对采集到的数据进行分析。为了有效利用传感数据中的信息,需要将分布在传感设备端的数据汇聚到服务器进行集中处理。传统的多址接入方案难以在短时间内传输海量数据,因此需要一种新型的快速数据汇聚方案。幸运的是,许多应用仅仅需要传感数据的统计信息(例如算数平均值、加权和等),因此服务器接收端无须复原所有的原始数据。基于这一特性,一种被称为空中计算

基金项目:国家重点研发计划(2019YFB1802804)、广东省基础与应用基础研究基金资助项目(2019B1515130003)

的新兴数据传输方式被业界提出,它能够利用信号在传输过程中的波形叠加属性,来实现快速的数据汇聚^[11]。

与传统多址接入方案不同,空中计算旨在降低收集到的统计信息与真实值之间的误差^[12]。这一误差往往通过均方差来衡量,并受设备端发射功率的影响^[13]。一方面,设备发射功率的增大将有助于克服噪声影响,从而降低均方差;另一方面,单独增大某几个设备的发射功率将会使各设备间的信号幅度差异过大,从而导致均方差增大。因此,需要对所有设备的发射功率进行统一调节,达到最优的空中计算性能。在无源的传感器网络中,可以采用无线供电的方式为空中计算供能,各设备的发射功率受限于其收到的能量^[14-15]。

为了实现超高速的数据处理,本文中我们提出一种基于空中计算的无线群智感知设计。该设计通过对无线功率分配策略、感知数据量,以及空中计算数据汇聚时间3个因素进

行联合优化,从而实现服务器数据开发效益的最大化。

1 基于频分复用的无线群智感知系统

如图1所示,本文中我们考虑由一个多天线服务器和 N 个单天线传感设备组成的多用户无线群智感知系统。服务器依据各用户反馈的信道状态和感知能力来调整无线功率传输策略,其中分配给各用户的功率 P_n 之和不得超过服务器的发射功率 P_0 ,即:

$$\sum_{n=1}^N P_n \leq P_0, \quad (1)$$

其中,给定无线功率传输时间为 T_0 ,能量转化效率为 η ,信道功率增益为 g_n ,各设备接收到的能量可以用 $E_n = \eta g_n P_n T_0$ 表示,并被划分为3个部分: $E_n^{(s)}$ 用于数据感知, $E_n^{(t)}$ 用于数据传输, $E_n^{(r)}$ 作为执行数据感知任务获得的能量奖励。

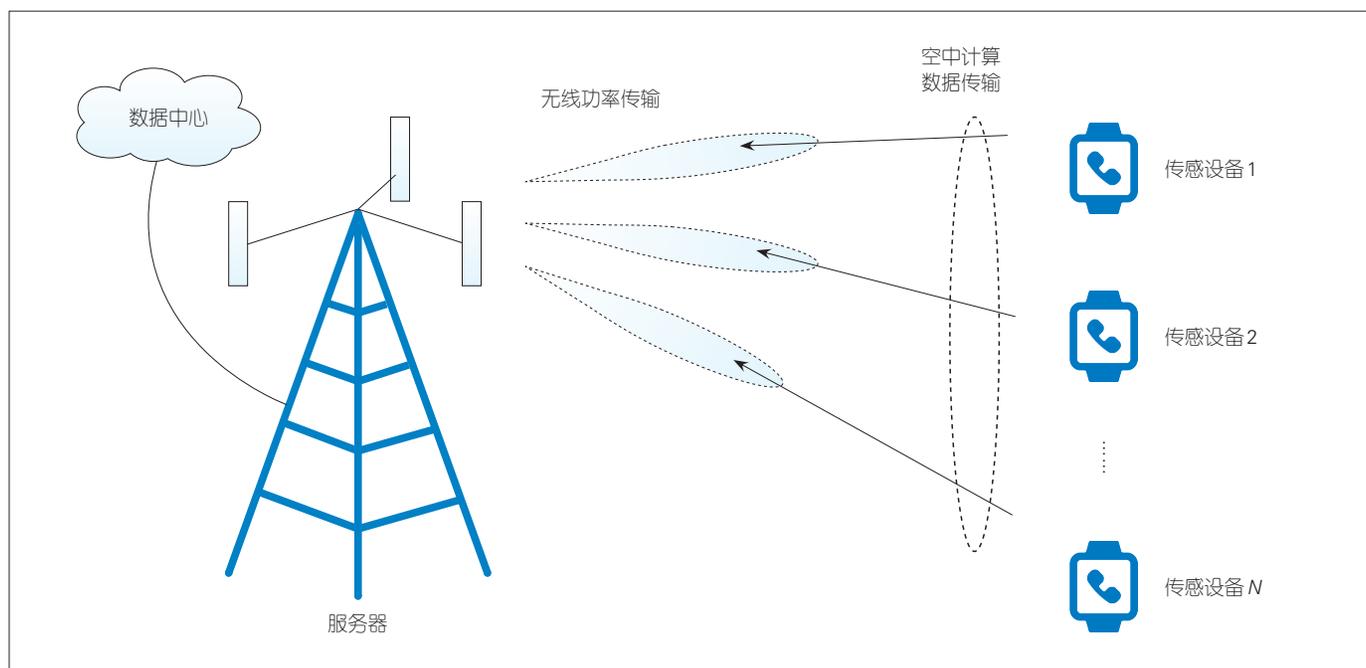
相应地,群智感知任务时间 T 也

将划分为两个部分,其中 t_n^s 用于数据感知, t_n^t 用于数据传输。因此,本文中我们考虑的设计需要同时满足能量和时间两个限制条件,如式(2)和(3):

$$E_n^{(s)} + E_n^{(t)} + E_n^{(r)} \leq E_n, \quad (2)$$

$$t_n^s + t_n^t \leq T. \quad (3)$$

对于数据感知而言,给定感知速率 s_n ,感知数据量可以用 $L_n = s_n t_n^s$ 来表示。相应地,给定感知单位数据量所需的能量 $q_n^{(s)}$,感知能量消耗可以用 $E_n^{(s)} = q_n^{(s)} L_n$ 来表示。同理,当感知单位数据的能量奖励 $q_n^{(r)}$ 确定时,可以得出各设备的数据感知能量奖励 $E_n^{(r)} = q_n^{(r)} L_n$ 。在传统的传输过程中,为了避免各设备感知数据在传输过程中的干扰,需要用频分复用的方式将各设备的信号在不同的频段发射。假设可用总带宽 B 被所有设备均分,每个设备分得的带宽为 B/M 。给定各设备的感知数据量、数据传输时间、



▲图1 基于空中计算的无线群智感知系统

信道状态,那么各设备的发射能量如式(4):

$$E_n^{(i)} = \frac{t_n \sigma^2}{g_n} \left(2^{\frac{M t_n}{B t_n}} - 1 \right), \quad (4)$$

其中, σ^2 表示噪声功率。基于文献[16],服务器数据开发效益 R 由感知数据效益和能量开销共同决定,即:

$$R = \sum_{n=1}^N a_n \log(1 + L_n) - c \sum_{n=1}^N P_n T_0, \quad (5)$$

其中, a_n 代表第 n 个设备的数据重要程度, c 代表单位能量开销的代价。为了最大化服务器数据开发效益,该系统将对无线功率分配、感知数据量,以及空中计算数据汇聚时间进行联合优化,优化问题构建如式(6):

$$\max_{\{P_n \geq 0\}, \{t_n \geq 0\}, \{L_n \geq 0\}} \sum_{n=1}^N a_n \log(1 + L_n) - c \sum_{n=1}^N P_n T_0, \quad (6a)$$

$$\text{s.t.} \sum_{n=1}^N P_n \leq P_0, \quad (6b)$$

$$(P1) L_n / s_n + t_n \leq T, \forall n, \quad (6c)$$

$$(q_n^{(s)} + q_n^{(r)}) L_n + \frac{t_n \sigma^2}{g_n} \left(2^{\frac{L_n M}{t_n B}} - 1 \right) \leq \eta g_n P_n T_0, \forall n, \quad (6d)$$

其中,第1个限制条件要求分配给各用户的功率之和不得超过服务器的发射功率;第2个限制条件要求各用户的群智感知任务需要在规定时间内完成;第3个限制条件要求各用户消耗的能量不得大于其接受到的能量。易证明当后两个限制条件取等号时该问题达到最优解,否则可以通过分配更多的时间或能量来提升感知数据量。因此,问题(P1)可以进一步简化为:

$$\max_{\{t_n\}} \sum_{n=1}^N a_n \log(1 + s_n(T - t_n)) - c \sum_{n=1}^N \left[\frac{(q_n^{(s)} + q_n^{(r)}) s_n (T - t_n)}{\eta g_n} + \frac{t_n \sigma^2}{\eta g_n^2} \left(2^{\frac{s_n M (T - t_n)}{t_n B}} - 1 \right) \right], \quad (7a)$$

$$\text{s.t.} \sum_{n=1}^N \left[\frac{(q_n^{(s)} + q_n^{(r)}) s_n (T - t_n)}{\eta g_n} + \frac{t_n \sigma^2}{\eta g_n^2} \left(2^{\frac{s_n M (T - t_n)}{t_n B}} - 1 \right) \right] \leq P_0 T_0, \quad (7b)$$

$$(P2) 0 \leq t_n \leq T, \forall n. \quad (7c)$$

对问题(P2)的目标函数求二阶导,易证明该问题为一个凸优化问题,其拉格朗日函数如式(8)所示:

$$L(t_n, \lambda) = (\lambda + c) \sum_{n=1}^N \left[\frac{(q_n^{(s)} + q_n^{(r)}) s_n (T - t_n)}{\eta g_n} + \frac{t_n \sigma^2}{\eta g_n^2} \left(2^{\frac{s_n M (T - t_n)}{t_n B}} - 1 \right) \right] - \sum_{n=1}^N a_n \log(1 + s_n(T - t_n)) - \lambda P_0 T_0. \quad (8)$$

对该函数应用昆恩塔克条件进行分析,可以得出该问题的最优解 $\{t_n^*\}$ 需要满足如式(9)的条件:

$$\frac{(\lambda^* + c) \sigma^2}{\eta g_n^2} \left[\left(1 - \frac{s_n M T \ln 2}{t_n^* B} \right) 2^{\frac{s_n M (T - t_n^*)}{t_n^* B}} - 1 \right] + \frac{s_n a_n}{1 + s_n (T - t_n^*)} - \frac{(\lambda^* + c) s_n (q_n^{(s)} + q_n^{(r)})}{\eta g_n} = 0, \quad (9)$$

其中, λ^* 为拉格朗日算子。基于该解可以得出最优的数据感知和功率分配策略,这两个策略均具备阈值结构,具体如式(10):

$$L_n^* = \begin{cases} s_n (T - t_n^*), & \varphi_n \geq \lambda^* \\ 0, & \varphi_n < \lambda^* \end{cases}, \quad (10)$$

$$P_n^* = \begin{cases} \frac{1}{\eta g_n T_0} \left[\frac{(q_n^{(s)} + q_n^{(r)}) s_n (T - t_n^*)}{\eta g_n} + \frac{t_n^* \sigma^2}{\eta g_n} \left(2^{\frac{s_n M (T - t_n^*)}{t_n^* B}} - 1 \right) \right], & \varphi_n \geq \lambda^* \\ 0, & \varphi_n < \lambda^* \end{cases}, \quad (11)$$

其中, $\varphi_n = a_n \eta g_n / (q_n^{(s)} + q_n^{(r)} + \sigma^2 M \ln 2 / g_n B)$ 为阈值判定函数。

2 基于空中计算的无线群智感知系统

对于服务器端仅需要感知数据统计信息的场景而言,各设备的感知数据在传输过程中直接进行波形叠加,因此无须避免各设备信号之间的干扰,可以采用空中计算的数据传输方案。每个用户均可用整个带宽 B 来传输数据,因此该问题的最优解 $\{t_n^*\}$ 需要满足如式(12):

$$\frac{(\lambda^* + c) \sigma^2}{\eta g_n^2} \left[\left(1 - \frac{s_n T \ln 2}{t_n^* B} \right) 2^{\frac{s_n (T - t_n^*)}{t_n^* B}} - 1 \right] + \frac{s_n a_n}{1 + s_n (T - t_n^*)} - \frac{(\lambda^* + c) s_n (q_n^{(s)} + q_n^{(r)})}{\eta g_n} = 0. \quad (12)$$

然而,由于各设备的感知数据必须同时传输才能实现正确的波形叠加,空中计算要求各设备的数据传输时间同步。为了解决这一问题,先完成某类型数据采集的设备需要等待其他设备采集完该类型数据后,才能同时开始数据传输,因此所有设备的传输时间为:

$$t^* = \min_n t_n^*. \quad (13)$$

相应的数据采集量和功率消耗可以表示为:

$$L_n^{*s} = \begin{cases} \min s_n (T - t^{*s}), \varphi_n' \geq \lambda^{*s} \\ 0, \varphi_n' < \lambda^{*s} \end{cases}, \quad (14)$$

$$P_n^{*s} = \begin{cases} \frac{1}{\eta g_n T_0} \left[(q_n^{(s)} + q_n^{(r)}) L_n^{*s} + \frac{t^{*s} \sigma^2}{\eta g_n} \left(2^{t^{*s}/B} - 1 \right) \right], \varphi_n' \geq \lambda^{*s} \\ 0, \varphi_n' < \lambda^{*s} \end{cases}, \quad (15)$$

其中， $\varphi_n' = a_n \eta g_n / (q_n^{(s)} + q_n^{(r)} + \sigma^2 \ln 2 / g_n B)$ 为阈值判定函数。

空中计算的引入将对无线群智感知系统的性能带来双面的影响：一方面，空中计算可以更加充分地利用有限的频谱资源提升数据传输速率，每个设备节省下来的时间可以采集更多的数据；另一方面，空中计算的同步性要求将造成部分设备在某时间段内空置，这段空置的等待时间会造成设备采集到的数据量下降。

3 仿真设计与分析

为了验证基于空中计算的无线群智感知设计性能，我们在 MATLAB 平台上进行了仿真验证。整个无线群智感知系统包括 1 个匹配 40 根天线的服务器和 10 个单天线传感设备，服务器与传感设备间的信道 g_n 服从莱斯分布。无线能量传输时间 T_0 和群智感知时间 T 均设置为 1 s，能量转化效率 η 设置为 0.5，总带宽 B 设置为 100 kHz，噪声功率 σ^2 设置为 10^{-9} W。对于每个传感设备而言，感知速率 s_n 服从 $[10^4, 10^5]$ bit/s 的均匀分布，感知单位数据量所需的能量 $q_n^{(s)}$ 服从 $[10^{-12}, 10^{-11}]$ J/bit 的均匀分布，感知单位数据量所获得的能量奖励 $q_n^{(r)}$ 服从 $[10^{-14}, 10^{-13}]$ J/bit 的均匀分布。

图 2 展示了服务器数据开发效益 R 随服务器的发射功率 P_0 的变化曲线。可以看到，随着服务器发射

功率的增大，服务器数据开发效益增加并逐步趋向恒定值。这是由于当发射功率较小时，能量成为限制感知数据量的主要因素；当发射功率足够大时，限制条件将不再是能量，而是时间等因素。此外，基于空中计算的无线群智感知设计性能优于传统的多址接入方案。这证明空中计算带来的传输速率提升将有效减少数据传输时间，从而使各设备增加的数据感知时间超过同步性造成的空置时间。

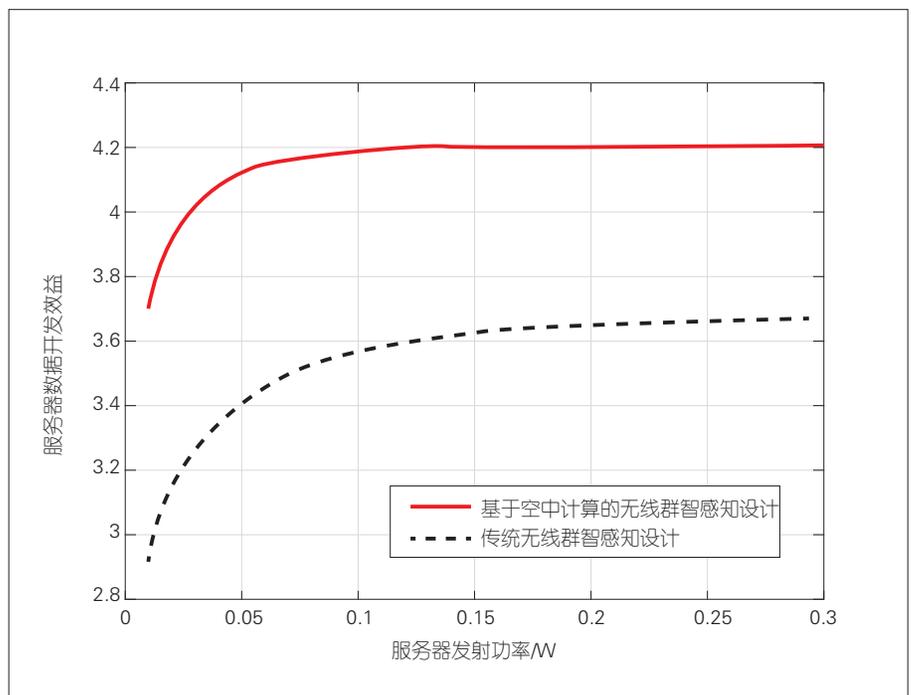
4 结束语

本文中，我们介绍了一种基于空中计算的无线群智感知系统。该系统通过对无线功率分配、感知数据量，以及空中计算数据汇聚时间进行联合优化，从而实现服务器数据开发效益的最大化。与传统多址接入方案相比，空中计算的引入有助于用户共享频谱资源，节省数据传输时间，从而增大数据感知量。然而，空中计

算在无线群智感知系统中的实现需要考虑更加实际的问题，比如需要通过功率控制来实现最小化空中计算均方差，需要对多天线空中计算进行波束赋形设计来实现多类型数据同时汇聚，这将成为未来的研究方向。

参考文献

- [1] AKYILDIZ I F, SU W, SANKARASUBRAMANIAM Y, et al. Wireless sensor networks: a survey [J]. Computer networks, 2002, 38(4): 393–422. DOI: 10.1016/s1389-1286(01)00302-4
- [2] GANTI R, YE F, LEI H. Mobile crowdsensing: current state and future challenges [J]. IEEE communications magazine, 2011, 49(11): 32–39. DOI: 10.1109/mcom.2011.6069707
- [3] LI X Y, ZHU G X, SHEN K M, et al. Joint annotator-and-spectrum allocation in wireless networks for crowd labelling [EB/OL]. (2019-12-25) [2020-05-28]. <https://arxiv.org/pdf/1912.11678.pdf>
- [4] LI X Y, ZHU G X, SHEN K M, et al. Spectrum allocation in wireless networks for crowd labelling [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Virtual Barcelona: IEEE, 2020: 8991–8995. DOI: 10.1109/icasspa40776.2020.9053492
- [5] ZHANG X, YANG Z, SUN W, et al. Incentives for mobile crowd sensing: a survey [J]. IEEE communications surveys and tutorials, 2016, 18(1): 54–67
- [6] LI X Y, YOU C S, ANDREEV S, et al. Wirelessly powered crowd sensing: joint power transfer, densing, compression, and trans-



▲图 2 空中计算与传统设计性能比较

- mission [J]. IEEE journal on selected areas in communications, 2019, 37(2): 391–406. DOI: 10.1109/jsac.2018.2872379
- [7] LI X Y, YOU C S, ANDREEV S, et al. Optimizing wirelessly powered over-the-air computation for high-mobility sensing [C]//IEEE Globecom Workshops(GC Wkshps). Abu Dhabi, UAE: IEEE, 2018: 1–6. DOI: 10.1109/IC-CW.2018.8403562
- [8] BI S Z, HO C K, ZHANG R. Wireless powered communication: opportunities and challenges [J]. IEEE communications magazine, 2015, 53(4): 117–125. DOI:10.1109/mcom.2015.7081084
- [9] LI X Y, HAN Z D, GONG Y. Adaptive multi-band resource allocation and power transmission system [C]//International Conference on Wireless Communications and Signal Processing(WCSP). Nanjing, China: WCSP, 2017: 1–6. DOI: 10.1109/WCSP.2017.8171174
- [10] CHEN M, HAN Z D, LI X Y, et al. Universal filtered multi-carrier based multi-user simultaneous wireless information and power transfer downlink system [C]//IEEE/CIC International Conference on Communications in China (ICCC). Beijing, China: IEEE, 2018: 410–415. DOI: 10.1109/ICCCChina.2018.8641173
- [11] ABARI O, RAHUL H, KATABI D, et al. Over-the-air function computation in sensor networks [EB/OL]. [2020–06–22]. <https://arxiv.org/abs/1612.02307>
- [12] ZHU G X, HUANG K B. MIMO over-the-air computation for high-mobility multimodal sensing [J]. IEEE Internet of Things journal, 2019, 6(4): 6089–6103. DOI: 10.1109/jiot.2018.2871070
- [13] CAO X W, ZHU G X, XU J, et al. Optimal power control for over-the-air computation in fading channels [EB/OL]. [2020–06–22]. <https://arxiv.org/abs/1906.06858>
- [14] LI X Y, ZHU G X, GONG Y, et al. Wirelessly powered data aggregation for iot via over-the-air function computation: beamforming and power control [J]. IEEE transactions on wireless communications, 2019, 18(7): 3437–3452. DOI:10.1109/twc.2019.2914046
- [15] LI X Y, ZHU G X, GONG Y, et al. Wirelessly powered over-the-air computation for high-mobility sensing [C]//IEEE Globecom Workshops(GC Wkshps). Abu Dhabi, UAE: IEEE, 2018: 1–6. DOI: 10.1109/GLO-COMW.2018.8644497
- [16] YANG D, XUE G, FANG X, et al. Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing [EB/OL]. [2020–06–22]. <https://www.semanticscholar.org/paper/Crowdsourcing-to-smartphones%3AAn-incentive-mechanism-Yang-Xue/4de1ca40f9bad26867601f4d6791bfbfe738653c>

作者简介



李晓阳, 南方科技大学在读博士; 主要研究领域为无线群智感知与群智标注、空中计算、无线功率传输等; 已发表论文10余篇。



贡毅, 南方科技大学电子与电气工程系教授, 2006–2018年先后担任《IEEE Transactions on Wireless Communications》和《IEEE Transactions on Vehicular Technology》的编委; 主要研究领域为5G与智能通信、移动边缘计算等; 承担多项国家级、省部级科技项目; 发表学术论文150余篇, 获得发明专利20余项。

← 上接第17页

用优化后的卷积神经网络来预测有墙环境的网络容量时, 可以在33 s内预测10 000个接入点部署方案的网络容量, 这比使用传统系统仿真方法所需要的12 413 s快376倍。

4 结束语

利用接入点部署的二维图像, 卷积神经网络将复杂环境中网络容量的预测转换为二维数据处理问题, 可成功提取接入点部署位置的特征, 实现高实时精准预测。比起传统的系统仿真方法, CNN更高效与智能, 且具有高精度和鲁棒性。随着人工智能技术的发展, 更多的机器学习方法将被应用于未来无线网络的部署与管理中。

参考文献

- [1] GUPTA A K, ANDREWS J G, HEATH R W. Macrodiversity in cellular networks with random blockages [J]. IEEE transactions on wireless communications, 2018, 17(2): 996–1010. DOI: 10.1109/TWC.2017.2773058
- [2] ONI P B, BLOSTEIN S D. Decentralized AP selection in large-scale wireless LANs considering multi-AP interference [C]//2017 International Conference on Computing, Networking and Communications (ICNC). USA: IEEE, 2017: 13–18. DOI: 10.1109/ICNC.2017.7876094
- [3] DEBNATH S, JEE A, BAISHYA S, et al. Access point planning for disaster scenario using dragonfly algorithm [C]//2018 5th International Conference on Signal Processing and Integrated Networks (SPIN). India, 2018: 226–231. DOI: 10.1109/SPIN.2018.8474051
- [4] RAN J, CHEN Y, LI S. Three-dimensional convolutional neural network based traffic classification for wireless communications [C]//2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP). USA: IEEE, 2018: 624–627. DOI: 10.1109/GlobalSIP.2018.8646659
- [5] MARSEET A, SAHIN F. Application of complex-valued convolutional neural network for next generation wireless networks [C]//2017 IEEE Western New York Image & Signal Processing Workshop (WNYISPW). USA: IEEE, 2017: 1–5. DOI: 10.1109/WNYIPW.2017.8356260
- [6] WANG X, WANG X, MAO S. CIFI: deep convolutional neural networks for indoor localization with 5 GHz Wi-Fi [C]//2017 IEEE International Conference on Communications (ICC). France, 2017: 1–6. DOI: 10.1109/ICC.2017.7997235
- [7] OROZA C A, ZHANG Z R, WATTEYNE T, et al. A machine-learning based connectivity model for complex terrain large-scale low-power wireless deployments [J]. IEEE transactions on cognitive communications and networking, 2017: 1–1. DOI: 10.1109/TCCN.2017.2741468
- [8] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述 [J]. 计算机学报, 2017, 40(6): 1229–1251. DOI: 10.11897/SP.J.1016.2017.01229
- [9] 严牧, 刘耀, 冯钢. 基于强化学习的无线网络智能接入控制技术 [J]. 中兴通讯技术, 2018, 46(2): 10–14. DOI: 10.3969/j.issn.1009-6868.2018.02.003
- [10] 张球, 盛敏, 李建东. 大数据驱动的“人工智能”无线网络 [J]. 中兴通讯技术, 2018, 24(2): 2–5. DOI: 10.3969/j.issn.1009-6868.2018.02.001

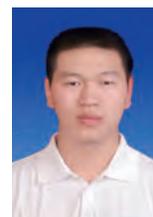
作者简介



赖昱辰, 华中科技大学在读硕士生; 主要研究方向为面向下一代无线通信技术与系统、人工智能技术等。



钟伟, 华中科技大学通信工程系讲师; 主要研究方向为异构蜂窝网、无线Ad-hoc网络、云无线接入网、CSMA及802.11 RTS/CTS机制、随机几何以及点过程理论等; 目前担任国际学术期刊《IEEE Wireless Communications Letters》《EURASIP Journal on Wireless Communication and Networking》《Physical Communication》编委。



王建峰, 微软公司研究员; 主要研究方向为机器学习、图像识别、物体检测、弱监督以及自监督特征学习等。



面向高效通信边缘学习网络的 通信计算一体化设计

Integrating Communication and Computation for
Communication-Efficient Edge Learning over Wireless Networks

朱光旭/ZHU Guangxu, 李航/LI Hang

(深圳市大数据研究院, 中国 深圳 518172)
(Shenzhen Research Institute of Big Data, Shenzhen 518172, China)

摘要:面向边缘学习网络,探讨了一种新型的基于空中计算的模型聚合方案,并对其中的关键使能技术展开论述。该方案利用无线多址信道的波形叠加特性将通信与计算在空中无缝融合,能够突破现有的通信-计算分离设计框架的局限性,从而大大提高频谱利用率,缓解了制约联邦式边缘学习大规模扩展的通信时延问题。

关键词:边缘智能;联邦式边缘学习;计算;多址接入

Abstract: A new model aggregation scheme based on over-the-air computing for edge learning over wireless networks is proposed, and the key enabling technologies are discussed. The proposed solution can achieve the desired model aggregation over the air via exploiting the wave-superposition property of multi-access channels, seamlessly integrating communication and computation. Therefore, it can break through the limitations of the classic design principle of decoupling communication and computation, greatly improve the spectrum efficiency and reduce the communication delay which restricts the large-scale expansion of federated edge learning.

Keywords: edge intelligence; federated edge learning; computation; multiple access

DOI: 10.12142/ZTETJ.202004006

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20200709.1041.002.html>

网络出版日期: 2020-07-09

收稿日期: 2020-06-11

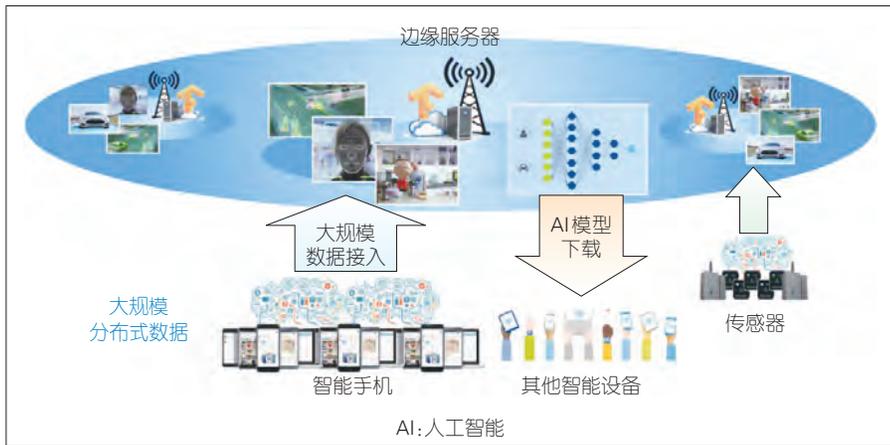
作为全新一代移动通信技术,5G将开启万物互联、深度融合的发展新阶段。受5G的推动,全球数据流量将呈现出爆炸式增长的趋势。据国际数据公司预测,到2025年,将有800亿台设备接入互联网,全球数据将达到163 ZB,是2016年数据的10倍。大数据的洪流加上不断取得

突破的人工智能(AI)技术激发了人们对泛在计算和智能的憧憬。在此愿景的推动下,越来越多的智能应用将被部署在网络边缘,而为之奠定基础边缘智能技术也正成为业界和学术界共同关注的热点^[1]。

边缘智能旨在为移动终端提供超快速、智能化和环境/位置感知的服务,这些服务包括虚拟现实(VR)/增强现实(AR)、自动驾驶、多媒体内容传输、智能家居和都市、工业自动化、电子银行、视频流分析。这些技术大多数都需要通过边缘机器学习来实

现,具体来说是将机器学习算法部署在网络边缘(如基站和智能终端)以快速地利用分布式的移动数据来连续地训练和调整边缘云(基站)中的人工智能模型(如图1所示)。机器学习所需的大量数据是由数百万到数十亿的物联网传感器和移动设备产生的。例如,谷歌公司为智能键盘而训练的人工智能模型要求数百万的移动设备同时上传用户交互数据。又如,特斯拉公司通过使用数百万特斯拉车辆在行驶过程中上传的雷达和激光雷达传感数据来不断改善其

基金项目:国家重点研发计划(2018YFB1800800)、广东省重点领域研发计划(2018B030338001)、广东省领军人才计划(00201501)、深圳市孔雀计划(KQTD2015033114415450)



▲图1 边缘智能系统

自动驾驶人工智能模型。由于数据量的巨大,将这些数据上传到边缘云给无线通信系统提出了极大的挑战。随着智能终端设备的爆炸式增长,多址接入延迟是实现低时延边缘机器学习的主要瓶颈。近10余年来,无线通信领域在空分多址(SDMA)、正交频分多址(OFDMA)和码分多址(CDMA)等多址接入技术方面取得了突破性进展。然而,对于这些正交接入技术而言,控制多址接入延迟需要无线资源随着设备数量线性增加。这意味着在有限频谱资源的限制下,正交多址接入技术难以扩展至大规模用户场景。

克服边缘机器学习的通信瓶颈

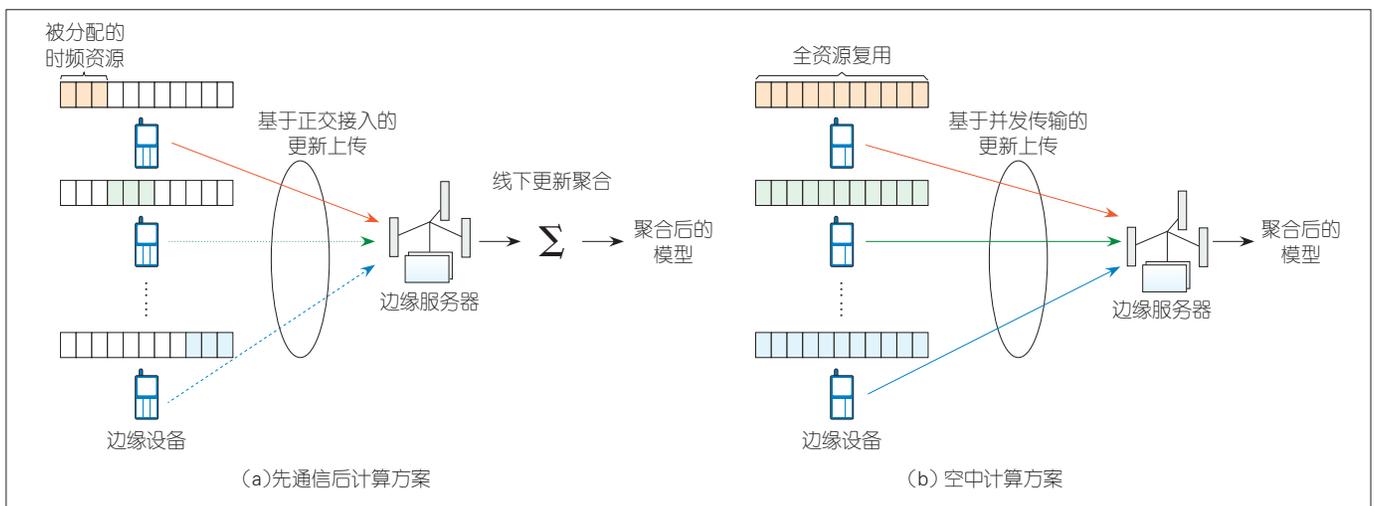
需要从根本上突破现有多址接入的设计原理和方法。正交多址接入技术是传统可靠通信设计理念的产物,其设计目标是在保证个体数据可靠传输的前提下最大化传输速率。在这一传统设计理念下,通信和计算是两个独立的过程:前者仅仅是为后者提供数据的“传输管道”。正是传统通信计算分离的设计思想造成了边缘机器学习中的通信瓶颈。这类技术无法对边缘学习系统的接收数据的后续计算应用进行整体考量,因而也无法进行更高效率的跨层优化。对于边缘学习系统来说,最终的任务是从海量数据中提炼出准确的AI模型,而非完成个体数据的可靠传输。

为了实现AI模型的准确训练,边缘服务器往往只对分布式数据所构成的某些特定函数而不是数据本身感兴趣。以目前最受关注的联邦式边缘学习为例,在其分布式模型训练的过程中,边缘服务器只需要从边缘设备上获取本地数据计算出的模型/梯度的平均值,而非所有存储在本地的个体数据^[2]。换言之,直接将基于个体数据可靠传输理念设计的通信技术套用到边缘学习系统中,将导致过低的频谱利用率以及不必要的通信时延。

综上所述,为了突破边缘学习中的通信瓶颈,亟待在计算与通信技术上进行改革与创新。为此,以计算和通信在空中的高度“融合”为特征的空中计算技术提供了一种解决之道。

1 空中计算概述

空中计算的概念最早起源于传感器网络中的数据聚合应用,其核心思想是利用无线多址接入信道的波形叠加特性以及多用户的并发传输以实现高速数据空中聚合^[3]。如图2所示,与传统的“先通信再计算”方案相比,它拥有极高的频谱利用效率,空中计算方案的接入时延不会随着



▲图2 先通信后计算方案对比空中计算方案

网络规模的增加而线性增加。早期的空中计算研究多聚焦于从信息论的角度分析其渐近计算性能。例如,在文献[4]中,基于高斯多址接入信道以及独立同分布的数据源假设,作者推导出了空中计算理论上的渐近计算速率,并证明了增加接入设备的数量能提高函数计算准确度这一令人鼓舞的结论。随后,无编码的模拟信号空中计算被证明可以在数据源服从单一高斯分布的情况下实现最小的计算误差^[5]。在空中计算的普适性方面,文献[6]的作者率先证明了通过适当的预处理和后处理,空中计算技术可以用于计算包括算术平均数和几何平均数在内的一系列被称为 nomographic 函数的统计函数。常见的 nomographic 函数如表 1 所示。在此基础上,文献[7]的作者进一步证明了任意函数都可以拆分为多个 nomographic 函数之和,这意味着空中计算具有处理任何函数计算的能力。这一里程碑式的发现大大拓宽了空中计算的应用场景。

空中计算令人满意的理论性能推动了一系列后续研究。这些研究

专注于解决实际信号处理问题和提高系统的鲁棒性,其中包括传感设备的功率控制设计^[8]、接入设备同步方案设计^[9],以及信道估计方案设计^[10]。除了在无线传感网络中的应用,空中计算中对并发干扰进行利用的思想也被广泛应用到现有的通信系统。例如,空中计算在中继系统上的应用催生了著名的计算转发中继方案,并大大提升了系统的抗噪性能^[11]。另一方面,通过空中计算中继的应用,始于有线网络的网络编码技术得以引入到无线网络中,并衍生出了广受关注的物理层网络编码研究领域^[12]。最新的一系列研究聚焦于利用多天

线技术及其带来的复用增益,将传统基于标量函数计算的空中计算技术拓展至矢量函数计算^[13-16],这对未来的多模传感网络数据聚合至关重要。

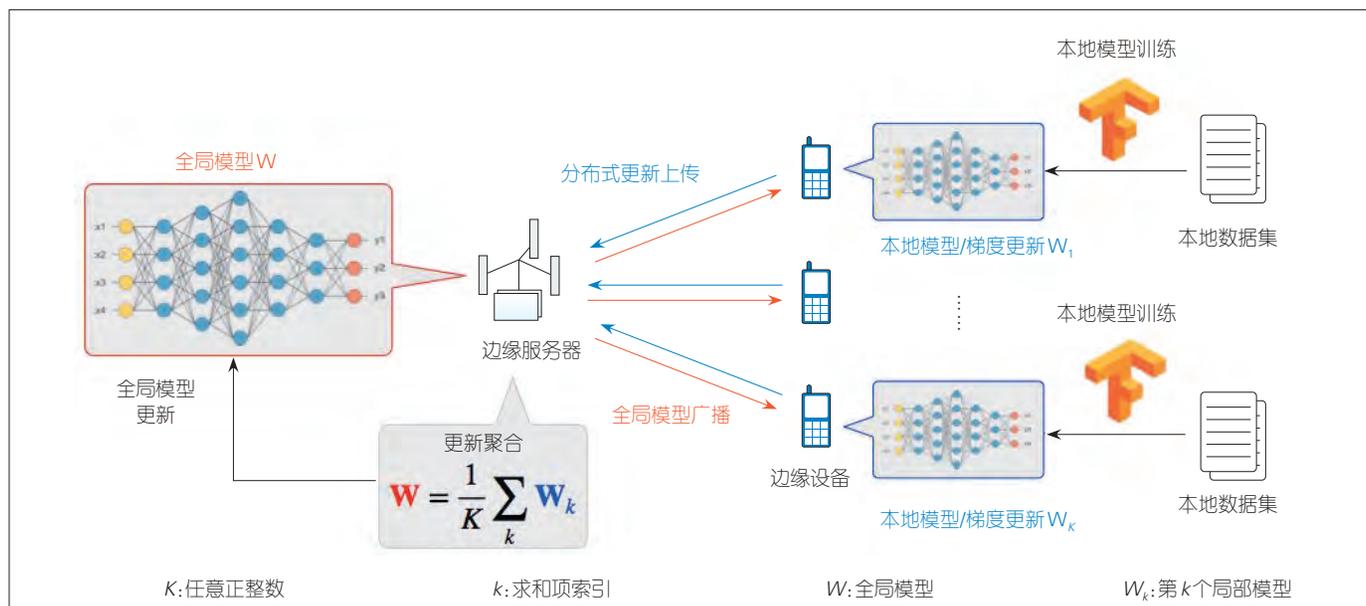
2 通信计算一体化设计关键技术方案

通过将差分隐私和分布式学习相结合,联邦式边缘学习成为目前最为流行的边缘学习范式。联邦式边缘学习的主要特点是将模型训练任务分配到参与训练的终端上,并以本地模型上传代替原始数据上传,这保障了数据隐私性的同时也有效利用了终端本地的计算资源。如图 3 所

▼表 1 常见的可用于空中计算的 nomographic 函数

名称	表达式
算术平均数	$y = \frac{1}{K} \sum_{k=1}^K x_k$
加权和	$y = \sum_{k=1}^K w_k x_k$
几何平均数	$y = (\prod_{k=1}^K x_k)^{1/K}$
多项式	$y = \sum_{k=1}^K w_k x_k^{\beta_k}$
欧几里得范数	$y = \sqrt{\sum_{k=1}^K x_k^2}$

β : 多项式中的指数项 k : 求和或连乘项索引 K : 任意正整数 w_k : 加权因子项



▲图 3 联邦式边缘学习

示,联邦式边缘学习中边缘服务器与边缘设备之间的交互在两个阶段之间交替进行。在第1阶段,边缘服务器将全局模型的当前版本广播给参与训练的边缘设备。基于当前广播模型,每个边缘设备使用随机梯度下降法并利用本地数据对本地模型进行更新。在第2阶段,边缘设备将其本地更新(梯度估计或模型更新)上传到边缘服务器,并进一步聚合以更新全局模型。这两个步骤的每次迭代称为一个通信回合,迭代一直持续到全局模型收敛。

鉴于模型/梯度更新的高维度性(一个典型的深度学习模型/梯度包含数百万至上亿个参数),由密集设备模型/梯度更新上传所带来的通信瓶颈是目前联邦式边缘学习所面临的一大挑战。现共有3种方法可以解决该问题:第1种方法是放弃响应速度慢的边缘设备的更新,以丢失部分更新信息为代价实现快速更新同步^[17];第2种方法是利用所提供模型/梯度更新的重要性而不是计算速度来对设备进行调度^[18-19];第3种方法致力于利用梯度更新的稀疏性^[20]和分辨率梯度参数量化^[21],来实现更新参数压缩。上述3种方法代表了现有研究通过“节流”的方式解决通信瓶颈问题,即通过设备调度策略和数据压缩以减少接入设备数和传输数据量,从而减轻通信负担。在这些方案中,无线信道仅被抽象为数据传输的“管道”,其特性并没有被充分利用来进行高效(模型/梯度)更新聚合方案的设计。另外,在联邦式边缘学习中,边缘服务器感兴趣的只是不同本地更新的平均值,而非本地更新自身的可靠传输,因此传统的基于个体数据可靠传输理论设计的正交多址接入技术将会带来不必要的通信延迟。为此,基于空中计算的更新聚合技术

应运而生,其高效的频谱利用率使其成为当前一大研究热点^[22-24]。目前已有初步的研究展示了基于空中计算的更新聚合技术在理论上的超低时延性能,然而该技术的落地仍然面临着不少实际的挑战。接下来我们将一一介绍其中的3大挑战,并对可能的解决方案进行论述。

2.1 空中计算的数字化和宽带化改造

传统针对无线传感网络数据聚合应用的空中计算方案主要面向窄带非频选信道,并且需要对发射信号进行高精度的模拟调制,即发射机可以根据需要调制载波波形,并自由选择同相/正交系数作为任意实数。然而现有的无线设备都带有嵌入式数字调制芯片,无法实现任意精度的模拟调制。另外,鉴于模型/梯度更新的高维度性,需要利用宽带信道对其进行传输,而由此产生的频选衰落将影响空中计算的精确度。因此,需要对传统窄带模拟空中计算技术进行数字化和宽带化的改造以实现联邦式边缘学习的通信计算一体化应用需求。

在现行的蜂窝系统中,主流的调制和宽带传输方案是正交幅度调制(QAM)和正交频分复用(OFDM)技术。为了更好地兼容现行的设备方案,我们提出一种基于QAM调制和OFDM架构的宽带数字空中计算系统方案。该新方案受符号随机梯度下降(signSGD)^[21]的启发,在边缘设备端进行1 bit信息量化,并在边缘服务器上进行基于多数表决的解码。下面我们基于梯度平均的联邦学习为例对该系统的收发机设计方案进行详细描述。

2.1.1 发射机方案

发射机方案如图4(a)所示。该设计建立在传统的OFDM发射机的

基础上,采用截断信道逆变功率控制。然而,与传统的通信系统需要先进行信道编码不同,我们将无编码的原始量化比特馈送到OFDM单元。特别地,我们先对边缘设备的梯度更新进行1 bit量化,即对每个梯度更新元素我们只取其符号位,如式(1):

$$\bar{g}_k = \text{sign}(g_k), \forall k, i, \quad (1)$$

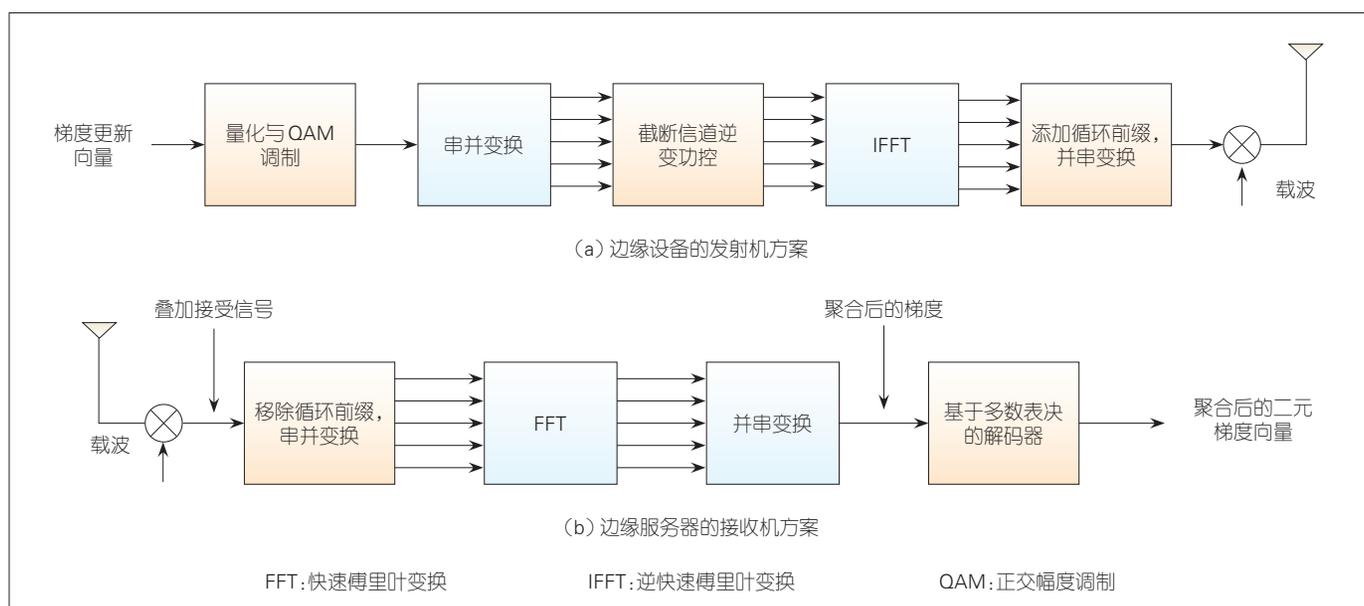
其中, g_k 是第 k 个设备的梯度更新矢量,而 \bar{g}_k 是其一比特量化后的对应矢量, $\text{sign}(\cdot)$ 代表的是取符号位操作。随后,每两个二进制梯度参数为一组被调制成一个4 QAM符号的同相和正交系数。假设OFDM系统共有 M 个子载波,调制后的长符号序列将被划分为块,每个块含有 M 个符号并作为单个OFDM符号被发送(其中每个频率子载波发送一个符号)。

假设发射机拥有完美的信道状态信息,我们可以通过信道逆变功率控制使得不同设备发送的梯度更新参数以相同的幅度被接受,从而实现空中计算所需的信道衰落对齐。为了在给定的功率约束下实现信道逆变,我们采用截断信道逆变功率策略,即只有当一个子信道的信道增益大于某一给定阈值时我们对信道逆变,否则我们将放弃使用该子信道(分配零功率)。

2.1.2 接收机方案

接收机方案如图4(b)所示。该方案具有与传统OFDM接收机相同的架构,只是数字检测器被替换为基于多数表决的解码器,用于根据接收信号估计全局梯度更新。

考虑一个任意的通信回合,假设所有的参与设备同步并发传输,并使用截断信道逆变功率策略,服务器所接收到的叠加的信号表示如式(2):



▲图4 数字化宽带空中计算的收发机设计

$$\bar{\mathbf{g}} = \sum_{k=1}^K \sqrt{\rho_0} \bar{\mathbf{g}}_k^{(T_r)} + \mathbf{z}, \quad (2)$$

其中, ρ_0 为采用截断信道逆变功控后的信道对齐水平, $\bar{\mathbf{g}}_k^{(T_r)}$ 是一比特量化梯度 $\bar{\mathbf{g}}_k$ 的“截断”版本, 其被截断的元素被置为零, 并取决于所属子信道的增益。最后, 为了从 $\bar{\mathbf{g}}$ 中获得全局梯度估计值用于模型更新, 我们采用基于多数表决的解码器如式(3):

$$\mathbf{v} = \text{sign}(\bar{\mathbf{g}}). \quad (3)$$

这里对叠加信号 $\bar{\mathbf{g}}$ 的每个元素取符号位的操作实际上实现了一个多数表决的判决机制, 即全局梯度的每个元素的符号由各局部梯度对应元素的符号按多数表决机制决定。服务器再通过将全局梯度估计广播给所有设备以进行本地模型更新, 随后新一轮通信回合被发起直至模型收敛。

在文献[25]中, 我们通过基于实际数据集的实验仿真, 对基于空中计算的更新聚合方案和传统基于正交频分多址接入的聚合方案进行性能比较。在模型训练准确度相当的前

提下, 与后者相比, 前者在时延上获得数10倍的降低, 展现了通信计算一体化设计的巨大潜力。

2.2 安全空中计算

在联邦式边缘学习中应用空中计算技术的时候, 需要关注的另一个实际问题是模型/梯度聚合的安全性。一些恶意用户可能通过故意上传不准确的模型/梯度更新或随机噪声来对模型训练进行攻击。这些恶意攻击可能导致学习算法无法收敛, 并使整个训练过程崩溃。为了保证模型训练的顺利进行, 需要建立安全可靠的计算机制以防范来自恶意用户的攻击。

为此, 一个可能的方案是采用直接序列扩频(简称扩频)技术对上传的模型/梯度更新进行编码。如图5所示, 在该方案中, 所有合法用户将使用由服务器分配的特殊扩频序列(也称作码片)进行扩频编码, 以保护其上传更新的合法性; 而不知道扩频序列的恶意用户所产生的攻击和干扰会在服务器的解扩频过程中得到有效抑制。以下是基于扩频技术的

空中计算方案的具体设计:

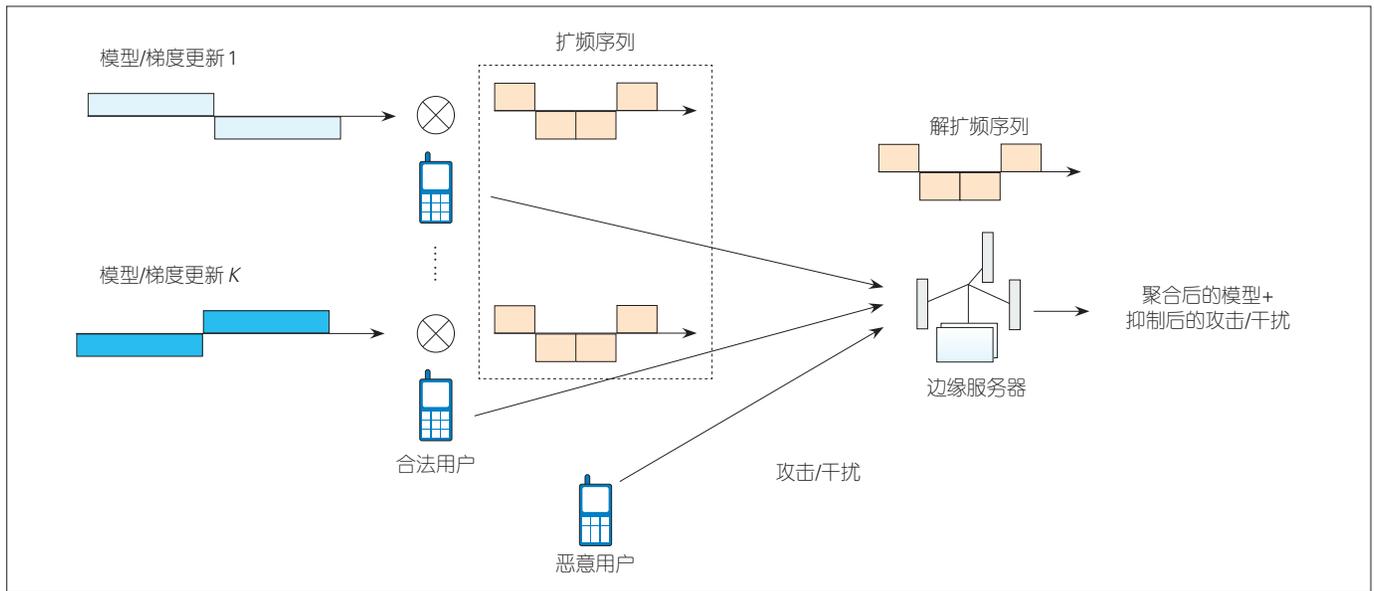
1) 网络中的所有合法用户都由服务器分配一个合法的扩频序列, 即一串取值+1或-1的伪随机噪声码序列, 恶意用户对该序列未知;

2) 所有合法用户将其模型/梯度更新信息与被分配的扩频序列相乘以进行扩频编码, 然后所有合法用户并发上传扩频后的更新信息;

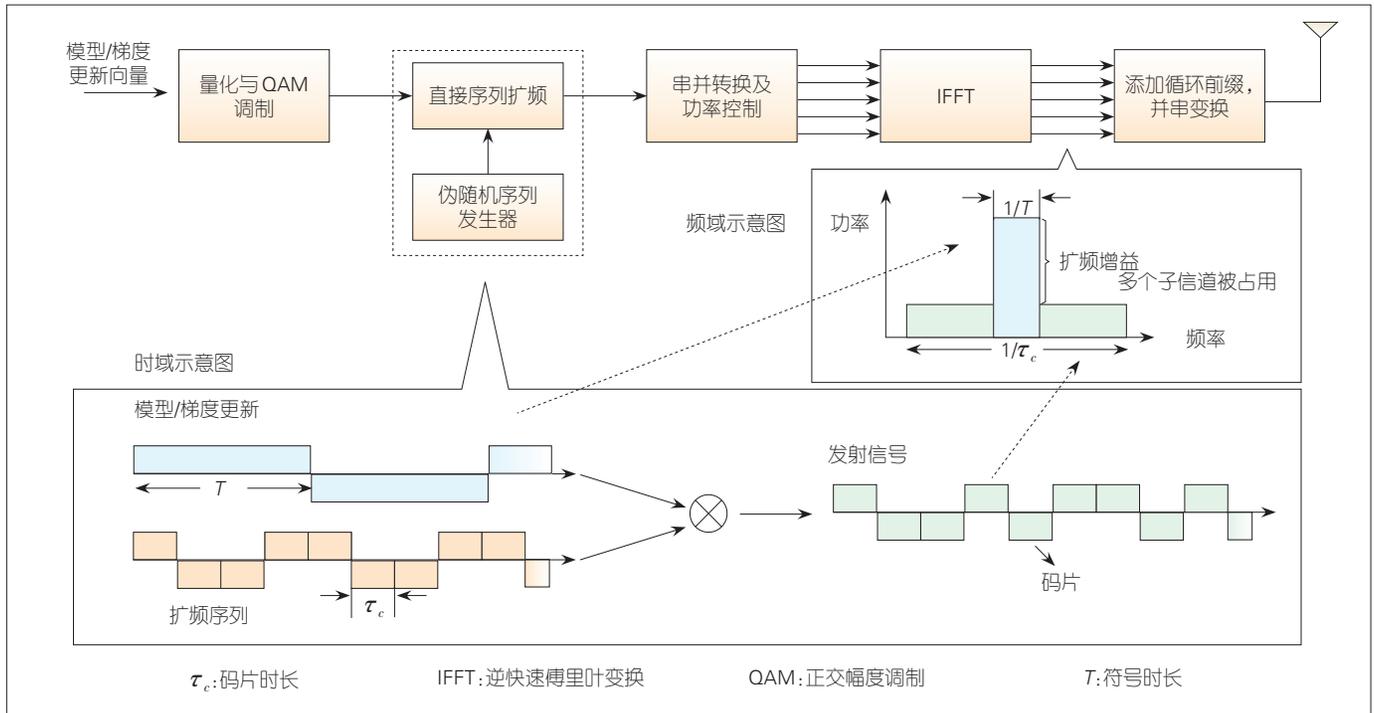
3) 最后, 边缘服务器使用约定的扩频序列, 对接收到的包含所有并发上传用户的叠加信号进行解扩频。

如图6所示, 基于扩频技术的空中计算的实现, 仅需要在图5(a)所示的OFDM框架上在调制模块和逆快速傅里叶变换(IFFT)模块中间添加一个额外的扩频模块即可。图6还展示了扩频技术背后的原理: 将上传更新信息和扩频序列相乘等价于用更大的带宽完成更新信息的上传。其中, 带宽扩展的倍数被称为扩频因子, 由符号时长 T 和码片时长 τ_c 的比率所决定, 即 $\gamma = T/\tau_c$ 。

所有合法用户使用同一个扩频序列进行编码, 因此边缘服务器对接收到的叠加信号进行解扩频后, 能够



▲图5 基于扩频技术的安全空中计算示意图



▲图6 基于扩频技术的安全空中计算的系统实现

自动聚合来自合法用户的模型/梯度更新,并同时以扩频因子的倍数抑制来自恶意用户的攻击。

值得一提的是,扩展因子的设计需要平衡时延代价和对恶意攻击抑制的强度之间的折中关系。具体而言,利用扩频技术后,系统在模型/梯度更新上传时占用的带宽是原来的 γ

倍。这在某种程度上削弱了空中计算相对于多用户正交传输所取得的时延优势。然而,以此代价换来的补偿为系统对恶意攻击的鲁棒性以及以扩频因子数倍升的信噪比。

2.3 弱信道受限的信道衰落对齐

基于空中计算的更新聚合方案

需要通过功率控制补偿不同用户的信道衰落,以满足空中计算所需的信道一致性,因此其性能会受限于小区边缘设备的弱信道。如果边缘服务器上装备有多天线阵列,则可以通过波束赋型设计来缓解小区边缘设备带来的性能瓶颈。该设计的核心思想是对这些小区边缘设备进行波束

聚焦,以补偿它们的路径损耗,从而提升小区边缘设备的信道质量。

值得注意的是,用于小区边缘设备信道增强的空中计算波束赋型在设计原理上不同于传统的空分多址波束赋型。空中计算波束赋型本质上是要尽量对齐不同设备到服务器之间的信道强度,从而利用并发“干扰”进行计算;而空分多址则试图利用波束赋型正交化多用户信道以抑制多用户间的串扰,以便来自不同用户的数据的可靠传输。具体的差异可以通过下面的波束赋型问题建模来进一步阐述。

我们考虑如下的一个多天线系统,一个装备有多天线的基站服务多个单天线的用户,其输入输出关系可表示如式(4):

$$y = \mathbf{F}^H \mathbf{H} \mathbf{x} + \mathbf{F}^H \mathbf{n}, \quad (4)$$

其中, $\mathbf{F} \in \mathbb{C}^{N \times K}$ 是待设计的波束赋型矩阵, N 代表装备在边缘服务器的天线数, K 代表边缘设备数。 $\mathbf{H} \in \mathbb{C}^{N \times K}$ 代表信道矩阵, 其中第 k 列代表的是第 k 个设备的信道向量。 $\mathbf{x} \in \mathbb{C}^K$ 代表发射信号向量, 其中的第 k 个元素代表第 k 个设备的发射信号。 \mathbf{n} 代表零均值的加性高斯白噪声向量, 其方差为 $E(\mathbf{n}^H \mathbf{n}) = N_0 \mathbf{I}$ 。

基于上述模型,用于小区边缘设备信道增强的空中计算波束赋型可以通过求解式(5)中的无约束最大化信噪比问题来设计:

$$(P1) \max_{\mathbf{F}} \frac{\text{Tr}(\mathbf{F}^H \bar{\mathbf{H}} \bar{\mathbf{H}}^H \mathbf{F})}{N_0 \text{Tr}(\mathbf{F}^H \mathbf{F})}, \quad (5)$$

其中,矩阵 $\bar{\mathbf{H}}$ 包含了有待增强的弱用户的信道向量。

而对于空分多址波束赋型,则要设计 K 个相互正交的波束向量,且每个都要在增强目标用户信道的同时迫零其余的干扰信道。我们将矩阵 \mathbf{F} 的第 k 列表示为 \mathbf{f}_k , 则空分多址的波束赋型设计可以建模为式(6)中的 K 个有约束最大化信道比问题:

$$(P2) \max_{\mathbf{f}_k} \frac{\mathbf{f}_k^H \mathbf{h}_k}{\sum_{g \neq k} \mathbf{f}_k^H \mathbf{h}_g + N_0} \quad (6)$$

$$\text{s.t. } \mathbf{f}_k^H \mathbf{h}_g = 0, \forall g \neq k.$$

通过比较式(5)和式(6)两个问题建模可发现,空分多址的实现需要天线数 N 大于用户数 K , 以确保有足够的空域自由度来满足信道正交约束,这对于拥有大量用户的大型网络来说并不可行。相比之下,空中计算波束赋型总是可行的,而更多的空域自由度可以用来增强弱用户的信

噪比。

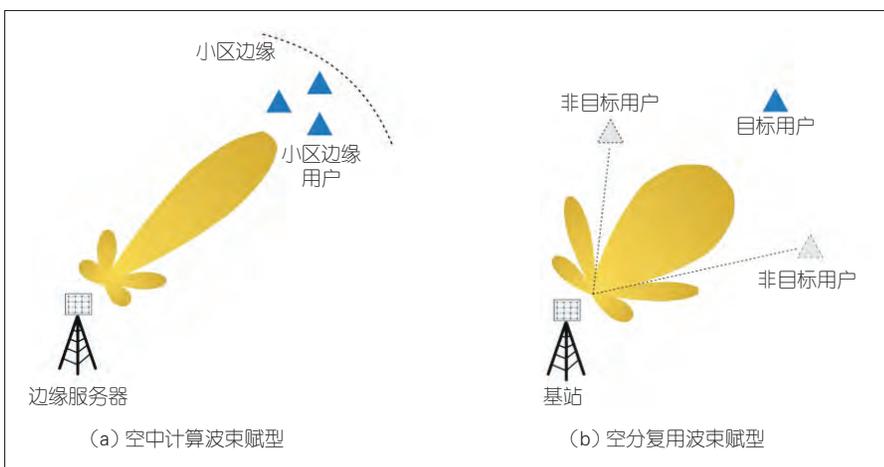
两种问题建模将导致截然不同的波束模式,具体如图7所示。一般来说,由于全空域自由度均用于信噪比增强,空中计算波束赋型可以形成更锐利和更强的波束以增强小区边缘用户的信噪比。与之相比,由于正交化约束消耗了相当的空域自由度,仅剩下部分自由度用于信噪比增强,空分多址的波束赋型对目标用户形成的波束相对较平坦,增幅较弱。此外,受限于天线阵列的空间分辨率,相邻(地理位置)的用户可能导致空分多址中的可分辨性问题。然而,由于空中计算波束赋型无需区分不同用户,并无此局限性。

3 结束语

在5G+AI的发展浪潮中,边缘学习是从分布在终端的海量数据中提炼AI的重要途径,也是将AI从云端推向网络边缘并实现泛在边缘智能愿景的重要技术;而通信时延瓶颈的解决是边缘学习向大规模用户场景扩展的关键突破。本文中我们所提倡的空中计算技术顺应了当前通信计算一体化的发展潮流,巧妙地利用并发传输造成的“干扰”进行快速数据聚合,大大提高了频谱利用效率并避免了计算中心对大量原始数据的存储,降低了大数据处理的负担。然而,高精度的可靠空中计算需要精确的信道估计、功率控制,以及设备间同步来支撑,因而如何提升空中计算在非完美条件下的鲁棒性是该技术走向成熟面临的关键问题。

参考文献

- [1] ZHU G X, LIU D Z, DU Y Q, et al. Toward an intelligent edge: wireless communication meets machine learning [J]. IEEE communi-



▲图7 空中计算波束赋型对比空分复用波束赋型

- cations magazine, 2020, 58(1): 19–25. DOI: 10.1109/mcom.001.1900103
- [2] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [EB/OL]. [2020–06–22]. <https://arxiv.org/abs/1602.05629>
- [3] 陈力, 卫国. 未来无线网络下的空中计算技术 [J]. 中兴通讯技术, 2019, 25(1): 29–34. DOI: 10.12142/ZTETJ.201901005
- [4] NAZER B, GASTPAR M. Computation over multiple-access channels [J]. IEEE transactions on information theory, 2007, 53(10): 3498–3516. DOI: 10.1109/tit.2007.904785
- [5] GASTPAR M. Uncoded transmission is exactly optimal for a simple gaussian “sensor” network [J]. IEEE transactions on information theory, 2008, 54(11): 5247–5251. DOI: 10.1109/tit.2008.929967
- [6] GOLDENBAUM M, BOCHE H, STANCZAK S. Harnessing interference for analog function computation in wireless sensor networks [J]. IEEE transactions on signal processing, 2013, 61(20): 4893–4906. DOI: 10.1109/tsp.2013.2272921
- [7] BUCK R C. Approximate complexity and functional representation [EB/OL]. [2020–06–22]. <https://www.sciencedirect.com/science/article/pii/S0022247X7990091X>
- [8] XIAO J J, CUI S, LUO Z Q, et al. Linear coherent decentralized estimation [J]. IEEE transactions signal processing, 2008, 56(2): 757–770. DOI: 10.1109/TSP.2007.906762
- [9] ABARI O, RAHUL H, KATABI D, et al. Air-Share: distributed coherent transmission made seamless [C]//2015 IEEE Conference on Computer Communications (INFOCOM). Hong Kong: IEEE, 2015: 1742–1750. DOI: 10.1109/infocom.2015.7218555
- [10] GOLDENBAUM M, STANCZAK S. On the channel estimation effort for analog computation over wireless multiple-access channels [J]. IEEE wireless communications letters, 2014, 3(3): 261–264. DOI: 10.1109/wcl.2014.022314.140005
- [11] NAZER B, GASTPAR M. Compute-and-forward: harnessing interference through structured codes [J]. IEEE transactions on information theory, 2011, 57(10): 6463–6486. DOI: 10.1109/tit.2011.2165816
- [12] NAZER B, GASTPAR M. Reliable physical layer network coding [J]. Proceedings of the IEEE, 2011, 99(3): 438–460. DOI: 10.1109/jproc.2010.2094170
- [13] ZHU G X, HUANG K B. MIMO Over-the-air computation for high-mobility multimodal sensing [J]. IEEE Internet of Things journal, 2019, 6(4): 6089–6103. DOI: 10.1109/jiot.2018.2871070
- [14] LI X Y, ZHU G X, GONG Y, et al. Wirelessly powered data aggregation for IoT via over-the-air function computation: beamforming and power control [J]. IEEE transactions on wireless communications, 2019, 18(7): 3437–3452. DOI: 10.1109/twc.2019.2914046
- [15] WEN D Z, ZHU G X, HUANG K B. Reduced-dimension design of MIMO over-the-air computing for data aggregation in clustered IoT networks [J]. IEEE transactions on wireless communications, 2019, 18(11): 5255–5268. DOI: 10.1109/twc.2019.2934956
- [16] CHEN L, ZHAO N, CHEN Y F, et al. Over-the-air computation for IoT networks: computing multiple functions with antenna arrays [J]. IEEE Internet of things journal, 2018, 5(6): 5296–5306. DOI: 10.1109/jiot.2018.2843321
- [17] TANDON R, LEI Q, DIMAKIS A G, et al. Gradient coding: avoiding stragglers in distributed learning [EB/OL]. [2020–06–22]. <http://proceedings.mlr.press/v70/tandon17a.html>
- [18] KAMP M, ADILOVA L, SINKING J, et al. Efficient decentralized deep learning by dynamic model averaging [EB/OL]. [2020–06–22]. https://link.springer.com/chapter/10.1007%2F978-3-030-10925-7_24
- [19] CHEN T Y, GIANNAKIS G, SUN T, et al. LAG: lazily aggregated gradient for communication-efficient distributed learning [EB/OL]. [2020–06–22]. <https://arxiv.org/abs/1805.09965>
- [20] LIN Y J, HAN S, MAO H Z, et al. Deep gradient compression: reducing the communication bandwidth for distributed training [EB/OL]. [2020–06–22]. <https://arxiv.org/abs/1712.01887>
- [21] BERNSTEIN J, WANG Y X, AZIZZAD-ENESHELI K K, et al. SignSGD: compressed optimization for non-convex problems [EB/OL]. [2020–06–22]. <https://arxiv.org/abs/1802.04434>
- [22] ZHU G, WANG Y, HUANG K. Broadband analog aggregation for low-latency federated edge learning [J]. IEEE transactions on wireless communications, 2020, 19(1): 491–506. DOI: 10.1109/TWC.2019.2946245
- [23] MOHAMMADI AMIRI M, GUNDUZ D. Machine learning at the wireless edge: distributed stochastic gradient descent over-the-air [J]. IEEE transactions on signal processing, 2020, 68: 2155–2169. DOI: 10.1109/tsp.2020.2981904
- [24] YANG K, JIANG T, SHI Y M, et al. Federated learning via over-the-air computation [J]. IEEE transactions on wireless communications, 2020, 19(3): 2022–2035. DOI: 10.1109/twc.2019.2961673
- [25] ZHU G, DU Y, GUNDUZ D, et al. One-bit over-the-air aggregation for communication-efficient federated edge learning: design and convergence analysis [EB/OL]. [2020–06–22]. <https://arxiv.org/pdf/2001.05713.pdf>

作者简介



朱光旭, 深圳市大数据研究院研究科学家; 主要从事无线通信理论研究, 包括智能通信、5G/B5G 通信技术等; 作为骨干成员参与国家重点研发计划及广东省重点领域项目, 曾获香港政府奖学金、国际会议 W C S P 最佳论文奖; 发表 S C I / E I 检索论文近 40 篇。



李航, 深圳市大数据研究院研究科学家; 主要研究方向包括无线网络、可见光通信、机器学习方法的应用; 曾担任 IEEE 多个会议的技术委员会委员, 多个专业期刊的审稿人; 发表论文 20 余篇。

面向边缘智能的空中计算

Over-the-Air Computation for Edge Intelligence



曹晓雯/CAO Xiaowen^{1,2}, 莫小鹏/MO Xiaopeng^{1,2}, 许杰/XU Jie¹

(1. 香港中文大学(深圳), 中国 深圳 440307;

2. 广东工业大学, 中国 广州 510006)

(1. The Chinese University of Hong Kong (Shenzhen), Shenzhen 440307, China;

2. Guangdong University of Technology, Guangzhou 510006, China)

摘要:边缘智能模型训练中的无线通信开销已成为系统性能瓶颈,空中计算是解决该问题的重要技术。利用无线多址接入信道的信号叠加特性,空中计算技术能够在多终端无线信号传输的同时,对参数汇总计算,从而实现通信计算一体化设计,降低无线通信开销,提高边缘智能系统性能。通过实例介绍了空中计算的基本原理及其在边缘智能中的应用,并展望了未来研究方向。

关键词:边缘智能;空中计算;分布式机器学习;多址接入信道

Abstract: The wireless communication overheads in the training of edge intelligence models have become the bottleneck of system performance. Over-the-air computation has emerged as a promising solution to address this issue. By exploiting the signal superposition property of wireless multiple access channels, over-the-air computation implements the aggregation of model parameters in a swift manner, during the concurrent transmission of multiple terminal devices. Via such an integrated communication and computation design, this technique can significantly reduce the wireless communication overhead and improve the AI training performance. Through case analysis, the basic principles of over-the-air computation and its application in edge intelligence are first introduced, and then the future research directions are presented.

Keywords: edge intelligence; over-the-air computation; distributed machine learning; multiple access channel

DOI: 10.12142/ZTETJ.202004007

网络出版地址: <https://kns.cnki.net/KCMS/detail/34.1228.TN.20200713.1123.002.html>

网络出版日期: 2020-07-13

收稿日期: 2020-06-08

1 边缘智能的概念

随着人工智能和物联网等技术的快速发展,信息技术与通信技术不断融合,通信网络已不再局限于提供数据传输服务,而正逐渐演变为

支撑下一代互联网、智能城市、自动驾驶、工业自动化的核心基础设施。近年来,通信网络支撑的终端设备数目和承载的业务量都急剧增加。据思科预测,到2023年,全球联网设备总数将达到293亿,而从2017—2022年,平均每年全球业务量增长将达到42%^[1];因此,未来通信网络需要支持海量设备节点的随时随地接入,并提供超可靠低时延的信息感知、传输、

处理、控制。

在通信网络中融入智能能力,是实现自动驾驶和工业自动化等新型应用的关键。未来的智能通信网络需要根据基站和终端设备所采集的海量数据进行学习和理解,进行智能的推理、规划和决策,并对物理世界进行反馈和执行控制。例如,在自动驾驶场景中,网络将汇总车辆的环境感知信息,并结合超高分辨率地图和实时交通信息,利用人工智能算法进

基金项目:广东省重点领域研发计划项目(2018B030338001)、国家重点研发计划项目(2018YFB1800800)、国家自然科学基金(61871137)、广东省普通高校省级科研项目(2018KZDXM028)、国家重点实验室开发研究基金项目(2019D08)

行智能推理和决策,辅助车辆进行导航路径规划和精准避障驾驶^[2]。由于海量数据产生于无线网络边缘,为实现快速的智能信息处理和实时控制,人们需要将传统云服务器的计算、存储和智能能力下沉到网络边缘的基站和终端设备;因此,边缘网络智能(或边缘智能)成为大势所趋,并成为未来6G研究的一个重要方向^[3-5]。与传统单机智能相比,边缘智能能够避免单个终端设备存储计算能力受限问题,打破设备间的数据孤岛;与传统云智能相比,边缘智能能够有效降低网络带宽需求,降低网络时延,保护数据隐私安全。

从技术上看,边缘智能主要包括在无线网络边缘对人工智能或机器学习模型进行分布式的智能训练和智能推理两个过程。其中,机器学习模型的智能训练对数据量和计算量有很高的要求,因此,本文中我们着重讨论机器学习模型的智能训练。在多种不同的分布式模型训练方法中,联邦学习^[6]在保障用户隐私和数据安全方面具有独特的优势,因此获得了非常广泛的关注。在联邦学习中,海量终端设备利用各自的本地数据,在边缘服务器的协调下,联合训练共同的机器学习模型。联邦学习的训练过程可以基于分布式梯度下降法迭代进行:在每一次迭代中,不同终端设备根据各自的本地数据,更新局部模型参数,并通过无线信道将各自的局部模型参数上传至边缘服务器进行模型汇总,以更新全局模型参数。上述步骤迭代进行,直至全局模型参数收敛。联邦学习能够在终端设备不进行原始数据共享的情况下,充分挖掘边缘网络蕴藏的分布式计算存储能力,进行高效的模型训练。

虽然联邦学习在边缘智能中具

有独特的优势,但是其频繁的模式参数传输汇总过程也带来了技术上的挑战:终端设备和边缘服务器之间的无线通信过程正在成为联邦学习训练速度等性能的瓶颈^[3]。已有研究工作从不同角度对该问题进行了研究,例如根据网络状态对上传的机器学习模型进行自适应压缩^[7]、优化局部更新和全局汇总的次数^[8],都是降低通信开销的有效方法;然而,已有研究往往采用传统的多址接入方法(如正交频分复用等),需要对各个终端设备的上传的模型参数单独进行解码。当终端设备数目很大以及训练迭代次数很多时,将出现巨大的无线通信资源开销问题,因此,如何从信息理论和通信理论的角度,寻求适用于联邦学习的新型多址接入方式是一个重要的问题。

空中计算是解决上述问题的一种有效技术^[9]。与传统多址接入方式对多用户数据单独解码、通信计算分离设计不同,空中计算技术可以利用无线链路上行多址接入信道的信号叠加特性,直接在空中进行计算,完成终端数据的快速汇总平均。空中计算技术通过通信和计算的一体化设计,可以有效降低分布式训练过程中的通信开销和时延,提高边缘智能网络和联邦学习的训练效率;因此,基于空中计算的联邦学习,已成为边缘智能的一个重要研究热点。

2 空中计算的基本原理

空中计算是指利用无线信号传输过程中的叠加特性,在空中实现对来自不同用户数据的函数计算。以下针对一个典型的多址接入信道,介绍空中计算的基本原理。如图1所示,系统包含 K 个终端设备和一个基站(或边缘服务器),令 \mathcal{K} 表示终端集合。假设每个终端设备 k 的本地信息

为 X_k ,而基站的目标是根据接收到的终端信号对 $\{X_k\}$ 进行计算。为简化讨论,设基站拟计算函数为 $\{X_k\}$ 的平均值,即:

$$\tilde{f} = \frac{1}{K} \left(\sum_{k \in \mathcal{K}} X_k \right)$$

在传统的多址接入方案中,基站对各个设备的数据 X_k 进行分别解码,再汇总平均。与之不同的是,在空中计算中,每个终端设备首先进行归一化操作 $g_k(\cdot)$,得到处理后的本地信息为 $x_k \triangleq g_k(X_k)$,然后发送信息 $\alpha_k x_k$,其中, α_k 表示终端设备 k 的发送系数。边缘服务器在接收到信号后,将直接检测以得到平均值 $f = \frac{1}{K} \sum_{k=1}^K x_k$,再通

过去归一化操作 $g_k^{-1}(\cdot)$ 获得有效信息 $\tilde{f} = g_k^{-1}(f)$;因此,这部分的难点在于如何有效恢复期望信号 f 。具体而言,令 h_k 表示终端设备 k 到基站的信道参数,则基站接收到的信号为 $y = \sum_{k=1}^K h_k \alpha_k x_k + z$,其中 z 为噪声。在接收到该信号后,基站通过降噪处理可以得到信息为 $\hat{f} = \frac{y}{K\sqrt{\eta}}$,其中 η 为降噪因子。

在系统没有噪声存在的理想情况下,通过设置 $\alpha_k = \frac{h_k^\dagger}{|h_k|^2}$ 以及 $\eta = 1$,接收信号直接变为基站的期望信号,即 $\hat{f} = f$,其中, \dagger 表示共轭运算;因此,在这种情况下,通过一次传输就可以实现 K 个终端设备数据的平均计算,大大提高系统的频谱利用率,降低传输时延。

在实际的空中计算过程中,系统受到无线信道的衰落特性以及接收机噪声的影响,而终端设备的发送功率也往往有限;因此,如何在实际系统约束下,设计发送信号和接收机算法,是有效恢复期望信号 f 的关键。在无线网络中,空中计算具体有两种实现方式:模拟和数字的空中计算。

对于模拟的空中计算,每个终端设备不需要对感知到的环境数据进行编码,只需要对原始数据进行预处理,紧接着通过无线信道发送到边缘服务器,并进行平均处理^[10-11]。针对模拟空中计算,一般使用计算失真率作为主要衡量指标,以有效衡量空中计算的链路性能。例如,计算均方误差(MSE)是一种有效衡量空中计算失真的指标,其定义为 $MSE = \mathbb{E}[(\hat{f} - f)^2]$ 。针对单天线加性高斯白噪声信道,本地信息 $\{X_k\}$ 独立的场景,计算均方误差可以表示为^[11]:

$$MSE = \frac{1}{K^2} \mathbb{E} \left[\left(\sum_{k \in \mathcal{K}} x_k \left(\frac{\sqrt{p_k} |h_k|}{\sqrt{\eta}} - 1 \right) + \frac{z}{\sqrt{\eta}} \right)^2 \right] = \frac{1}{K^2} \left(\sum_{k \in \mathcal{K}} \left(\frac{\sqrt{p_k} |h_k|}{\sqrt{\eta}} - 1 \right)^2 + \frac{\sigma^2}{\eta} \right), \quad (1)$$

公式(1)中, σ^2 为噪声功率。在这种情况下,可以利用发送端的自适应功率分配,平衡信号的不对准以及

噪声的影响,有效降低系统均方误差^[11]。除此之外,为进一步提高空中计算性能,文献[12-14]采用多天线技术,利用多天线的空间复用和阵列增益,联合设计发送端和接收端波束赋形,可以同时实现多模态传感信息的矢量值函数计算,最大程度地降低均方误差。

在实际情况下,噪声会严重损害模拟空中计算的性能,较低的信噪比将导致严重的计算失真。数字空中计算通过终端设备的信源编码,是一种有效抑制噪声的手段^[15-16]。对于数字空中计算系统,计算速率是有效衡量链路性能的技术指标。计算速率 R 定义为当信号失真或解码误差趋向于无穷小情况下,每次信道实现系统能够计算的函数数量,即 $R = \frac{F}{n}$, 其中, F 为计算的函数数量, n 为信道实现总次数。例如,针对采用格形编码的场景,在加性高斯白噪声信道下,系统的可达计算速率(比特每信道实

现)可以描述为^[15]:

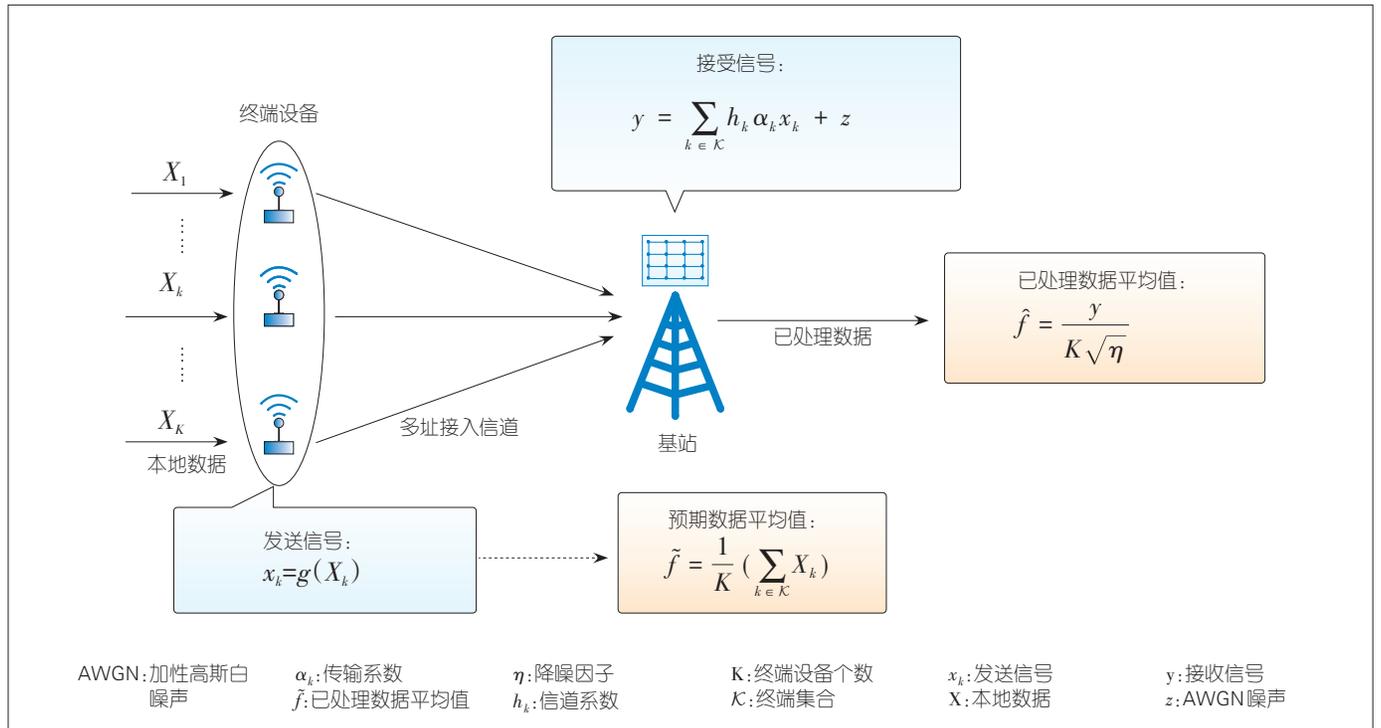
$$R = \left(\log \left(\frac{1}{K} + \min_{k \in \mathcal{K}} \frac{|h_k|^2 p_k}{\sigma^2} \right) \right)^+ \quad (2)$$

公式(2)中, $(x)^+ = \max(x, 0)$ 。针对不同场景,功率分配、多天线、非正交多址接入技术^[16]都是提高计算速率的有效手段。

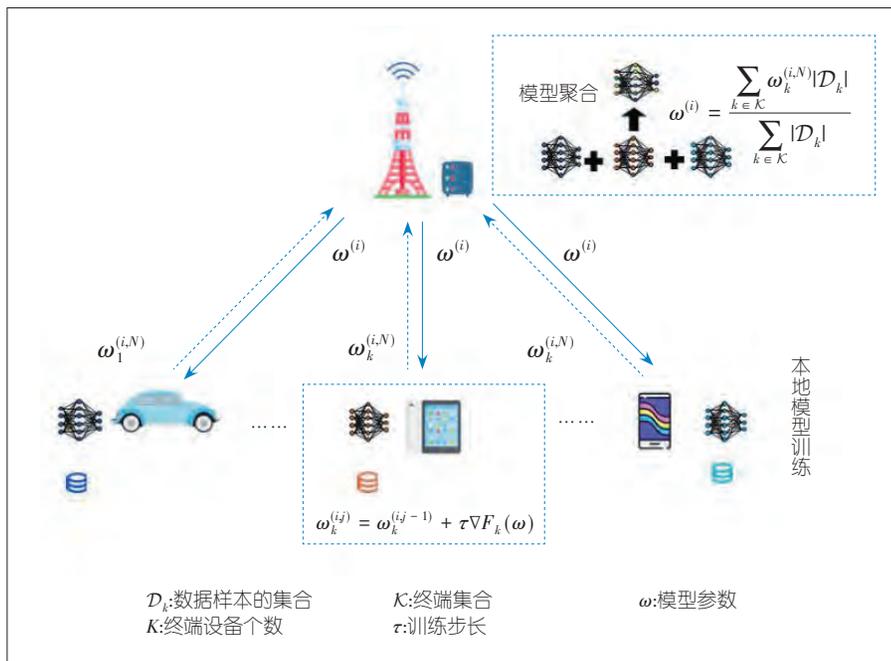
3 空中计算驱动的边缘智能

边缘智能是指通过挖掘基站和终端设备散布的数据和通信计算资源,在边缘网络中对人工智能或机器学习模型进行分布式的训练和推演。本节中我们主要关注联邦学习的分布式训练过程^[6]。如图2所示,系统中的 K 个终端设备利用各自的本地数据,在边缘服务器的协调下,联合训练一个共同的机器学习模型。

在该系统中,拟训练的机器学习模型参数表示为 ω , 每个终端设备 $k \in \mathcal{K}$ 的数据样本的集合表示为 \mathcal{D}_k ,



▲图1 空中计算示意图



▲图2 联邦学习示意图

该数据集 \mathcal{D}_k 的大小表示为 $|\mathcal{D}_k|$ 。假设 $f_i(\omega)$ 是数据样本 $i \in \mathcal{D}_k$ 的损失函数。例如,对于线性回归而言,假设每个样本的输入为 \mathbf{x}_i ,经过模型后输出为 y_i ,则对应的损失函数为 $f_i(\omega) = \frac{1}{2} |y_i - \omega^T \mathbf{x}_i|^2$ 。对于支持向量机而言,每个样本的损失函数为 $f_i(\omega) = \frac{\lambda}{2} |\omega|^2 + \frac{1}{2} \max(0, 1 - y_i \omega^T \mathbf{x}_i)^2$, λ 为常数。相应地,终端设备 $k \in \mathcal{K}$ 的本地平均损失函数就可以写为 $F_k(\omega) \triangleq \frac{1}{|\mathcal{D}_k|} \sum_{i \in \mathcal{D}_k} f_i(\omega)$ 。所有 K 个终端设备的全球平均损失函数为 $F(\omega) = \frac{\sum_{k=1}^K |\mathcal{D}_k| F_k(\omega)}{\sum_{k=1}^K |\mathcal{D}_k|}$ 。联邦学习的目标是通过批量梯度下降或随机梯度下降等梯度下降^[17]优化方法,找到期望的模型参数 ω ,以最小化全局平均损失函数 $F(\omega)$ 。

以分布式批量梯度下降算法为例,该训练过程将迭代进行。假设 M 代表边缘服务器的全局模型更新迭

代次数, N 代表终端设备本地模型迭代更新次数。对边缘服务器, $\omega^{(0)}$ 表示的初始全局模型参数, $\omega^{(i)}$ 表示第 i 次全局模型更新后的参数;对终端设备 k , $\omega_k^{(i,j)}$ 表示第 i 次全局模型更新中第 j 次本地模型更新的本地模型参数。具体而言,第 i 次全局模型更新的训练过程按照如下方式进行:

步骤(1):边缘服务器将全局模型参数 $\omega^{(i-1)}$ 广播下发至 K 个终端设备,同时终端设备的本地模型参数也相应同步设置为 $\omega^{(i-1)}$,即 $\omega_k^{(i,0)} = \omega^{(i-1)}$;

步骤(2):终端设备以迭代的方式,通过梯度下降的方法最小化本地平均损失函数 $F_k(\omega)$,以更新本地模型参数。假设 $\nabla F_k(\omega)$ 为本地平均损失函数 $F_k(\omega)$ 的梯度,则有 $\omega_k^{(i,j)} = \omega_k^{(i,j-1)} + \tau \nabla F_k(\omega)$,其中 τ 为训练步长。该本地更新过程将迭代 N 次;

步骤(3):终端设备将更新后的本地模型参数 $\omega_1^{(i,N)}, \dots, \omega_k^{(i,N)}$ 通过无线信道上传到边缘服务器;

步骤(4):边缘服务器根据接受

到的本地模型参数,通过加权平均来更新全局模型参数 $\omega^{(i)} = \frac{\sum_{k \in \mathcal{K}} \omega_k^{(i,N)} |\mathcal{D}_k|}{\sum_{k \in \mathcal{K}} |\mathcal{D}_k|}$ 。

经过 M 次全局模型迭代更新后,边缘服务器上的全局模型参数 $\omega^{(M)}$ 可作为所需的最小化 $F(\omega)$ 的解,即 $\omega^* \leftarrow \omega^{(M)}$ 。根据上述优化过程可知,基于传统的多址接入方式,边缘服务器在步骤(3)需要对不同终端的本地模型参数 $\omega_1^{(i,N)}, \dots, \omega_k^{(i,N)}$ 进行分别解码,再进行步骤(4)的汇总平均。当终端数目很大的时候,步骤(3)的通信过程将成为系统性能的瓶颈。因此,利用第2节介绍的空中计算计算,将步骤(3)和步骤(4)的通信和计算过程进行一体化设计,将能够很好地提高模型训练的性能。

近年来,利用空中计算进行高效的模型参数传输聚合,已成为解决联邦学习的通信瓶颈的一个研究热门^[18-22]。文献[18-22]研究了基于(模拟的)空中计算技术的边缘联邦学习,充分利用多址接入信道的信号叠加特性,提高联邦学习的收敛速度和准确度。由于在信息汇总过程中,信道较差的终端上传的模型参数将会产生较大的失真,影响网络整体的模型训练收敛速度。针对此问题,文献[18-22]分别从用户调度和功率控制的角度进行了研究。例如,文献[18]结合用户筛选和接收端波束赋形设计,在满足计算均方误差要求的情况下,最大化参与联邦学习的终端设备数目,以提高模型训练的性能;在功率控制方面,文献[19]针对一个宽带正交频分多址接入系统,提出一种截断功率控制方法,排除遭受深度衰落信道的终端设备,在学习性能和聚合误差之间取得良好的折衷;文献[20]研究了联邦学习中基于梯度统计信息的空中计算功率控制问题,关注衰

落信道下的最优功率控制问题,通过联合设计发射端的功率控制和接收端的降噪处理,最小化计算误差,进而提高收敛速度;此外,针对模拟空中计算中出现的噪声干扰问题,文献[21]研究了联邦学习中由于采用空中计算进行信息汇总出现的迭代噪声问题,依据子空间学习与跟踪技术解决了存在的传输数据缺失情况;文献[22]研究了基于空中计算的联邦学习出现的隐私保护问题,为了防止边缘服务器等聚合中心恶意揣测用户的隐私数据,通过对终端用户的功率进行控制,进而控制接收机接收到的注入到在聚合的全局模型中的噪声干扰,从而实现对保密级别和信噪比之间的最佳权衡。

4 未来研究展望

空中计算驱动的边缘智能具有巨大的应用前景,但其研究还处于初始阶段。例如,如何将数字空中计算与边缘智能进行有效结合?如何建立准确的性能度量体系,刻画空中计算下分布式边缘智能网络的性能极限?这些都还需要研究。除此之外,本节中我们对空中计算驱动的边缘智能的几个未来研究方向进行展望。

1) 分层网络的空中计算

已有工作主要研究单个边缘服务器和多个终端设备协作进行机器学习模型训练的场景。为充分挖掘边缘智能的潜力,需要利用大规模网络中海量终端设备的分布数据进行学习。在这种情况下,单个边缘服务器可能无法满足海量设备连接和计算能力的要求,因此,人们需要设计新的分层网络架构,通过依靠多个边缘服务器甚至云服务器,实现海量节点的分布式数据聚合和模型训练。针对分层网络,空中计算是提高分布式模型训练性能的一种有效手段。

以图3的三层网络为例,每个边缘服务器连接了不同的终端设备,而不同边缘服务器连接到上层服务器进行数据和模型的汇总更新。该三层网络可以通过两跳的空中计算实现大规模终端设备的模型汇总平均:在模型训练过程中,不同终端可以利用第一跳的空中计算,将更新后的局部模型参数上传至中间边缘服务器;边缘服务器则进行第二跳的空中计算,将其部分汇总的模型参数上传至上层服务器,进行全局模型汇总聚合。在这种情况下,如何确定中间边缘服务器的转发策略(基于模拟空中计算的放大转发或基于数字空中计算的解码转发)?如何确保两跳空中计算的时间同步?如何抑制或利用终端设备到不同中间边缘服务器的共道干扰?这些都是值得深入探索的问题。

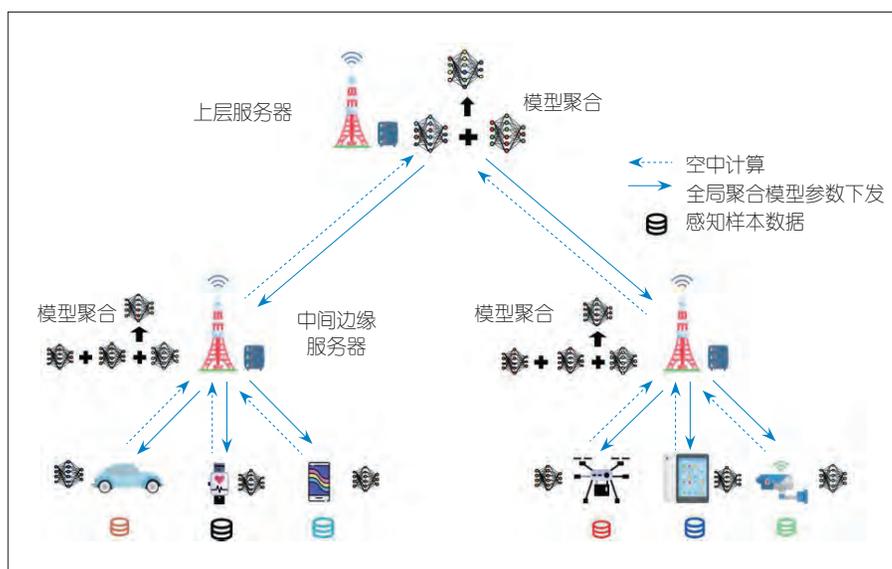
2) 设备间通信辅助的多跳空中计算

未来网络将存在海量终端设备,其中部分设备可能距离边缘服务器较远,这将大大影响空中计算和边缘智能的性能。为解决该问题,可以采用设备间通信技术,利用附近的一些闲置终端设备帮助偏远的设备终端

进行信息汇总。例如,通过在网络中将终端设备划分为多个簇,每个簇具有一个簇头和多个邻近的簇成员。由于相同簇中的终端设备之间距离很近,利用设备间通信能够保证较好的传输性能。如图4所示,簇成员通过将各自信息经由空中计算汇总到簇头端。所有簇头作为一个信息汇总中继节点,将接收到的信息通过空中计算汇总到边缘服务器。如何根据网络规模,设计分簇大小并选择适当的簇头节点?如何基于分簇情况,联合优化通信和计算资源,并设计设备间通信与空中计算技术的融合机制?这些都是未来重要的研究方向。此外,已有研究表明,通过在有限的信任簇群中通过设备间通信传输对数据进行共享,能有效改善分布式机器学习面临的数据非独立同分布问题^[23];因此,针对一些需要进行数据共享的边缘智能应用场景,如何有效结合设备间通信以及数据分布重塑增益,提升设备间通信辅助的空中计算性能,也是未来值得研究的方向。

3) 隐私保护

在边缘智能网络中,移动终端训练的模型数据需要通过无线信道发



▲图3 三层网络下的空中计算

送到边缘服务器进行聚合汇总。尽管在联邦学习中,终端设备不需要将其私有数据公开,但其仍然面临隐私泄露的风险。这是由于终端设备上上传的模型参数仍然存在有用信息,边缘服务器(或环境中的窃听器)可以从接收到的每个终端设备发送的信息中恶意地推断出终端设备的私有信息(例如数据的标签等)。空中计算则可以利用无线信道的叠加特性避免这一方面的隐私泄露。当网络中的终端设备将信息发送到空中后,边缘服务器接收到的是所有信息的叠加信号,无法从中推断出具体某一个终端设备的信息,从而避免终端设备的隐私被恶意推测;因此,在这种情况下,如何通过配置最优无线资源分配实现保密性高的空中计算?如何刻画联邦学习训练的收敛性、无线资源优化和隐私保护之间的最优折衷关系?这些都是值得深入探讨的关键问题。

4) 能量效率问题

在未来大规模物联网应用中,终端设备尤其是低功耗物联网节点的能量效率问题,显得至关重要。针对空中计算驱动的边缘智能网络,如何提高系统的能量效率是一个重要的问题。例如,联合优化终端的计算和通信资源分配是一个有效的方案。此外,也可以利用先进的能量技术(如能量采集和无线能量传输等),从能量供给侧提高边缘智能网络的能量效率和成本效益^[24]。例如,图5给出了一个可持续边缘智能网络的示意图,其中,边缘服务器利用环境中的可再生能源(比如太阳能、风能等)进行供能,而终端设备则利用无线能量传输供能,或利用无线反射通信进行模型上传。由于可再生能源到达具有随机性和间歇性,而无线能量传输效率则取决于发射端的功率和距

离,因此,如何联合优化能量管理和空中计算的无线资源是一个值得深入研究的课题。

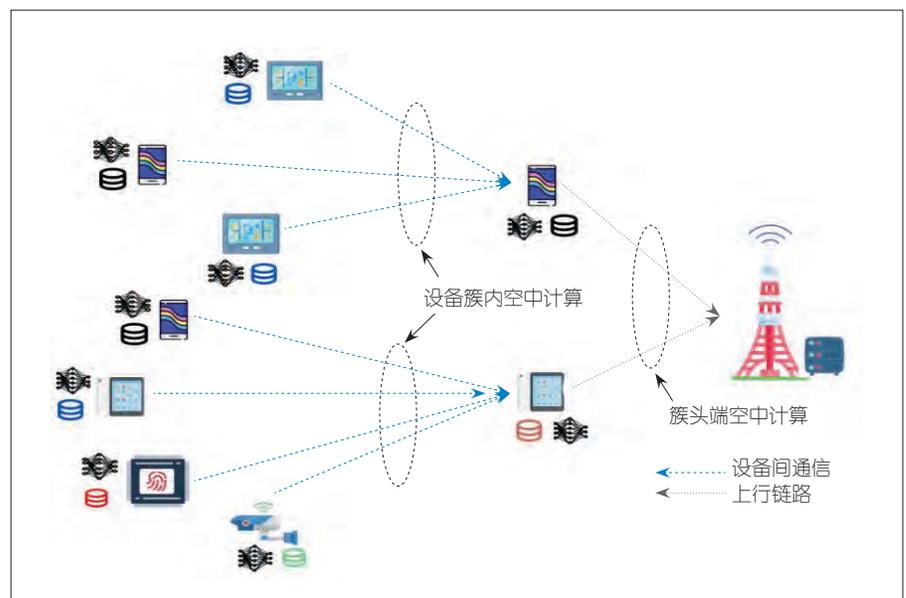
5 结束语

空中计算打破了传统无线网络通信计算分离的架构,实现“通信计算一体化”,能有效降低边缘智能网络的通信计算开销,进而提高训练性能。目前,针对空中计算驱动的边缘智能研究尚处于起步阶段。针对分

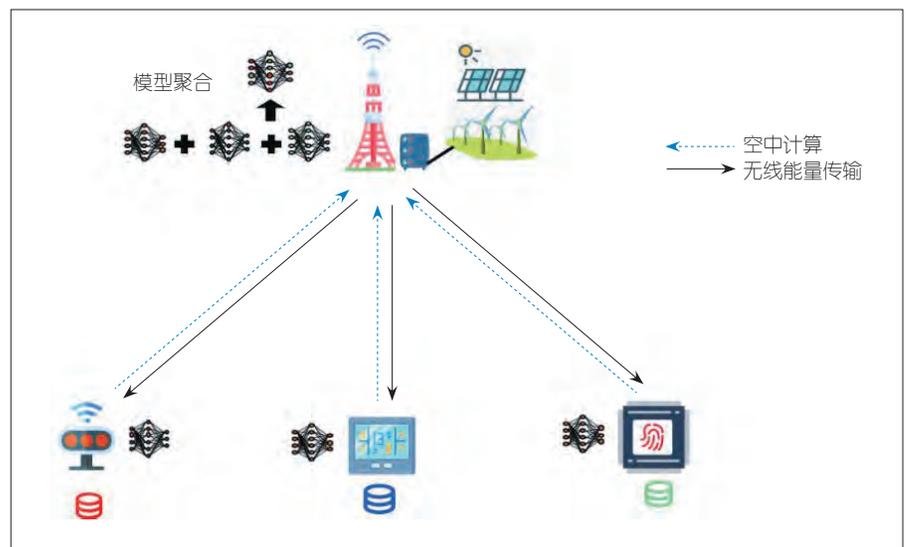
层网络和设备间通信等不同网络架构,考虑隐私保护和能量效率等实际因素,研究先进的空中计算方法,刻画边缘智能的性能极限,是未来研究亟待深入探讨的理论问题。这对推动空中计算走向6G应用具有重要的实际价值。

参考文献

- [1] Cisco. Cisco global cloud index: forecast and methodology, 2016–2021, white paper [EB/OL]. [2020-06-28]. <https://www.cisco.com/c>



▲图4 设备间通信辅助的空中计算



▲图5 可持续边缘智能网络示意图

- en/us/solutions/collateral/service-provider/globalcloud-index-gci/white-paper-c11-738085.html
- [2] LI E, ZENG L K, ZHOU Z, et al. Edge AI: on-demand accelerating deep neural network inference via edge computing [J]. IEEE transaction on wireless communication, 2020, 19(1): 447-457. DOI: 10.1109/TWC.2019.2946140
- [3] ZHU G X, LIU D Z, DU Y Q, et al. Toward an intelligent edge: wireless communication meets machine learning [J]. IEEE communication magazine, 2020, 58(1): 19-25. DOI: 10.1109/MCOM.001.1900103
- [4] LETAIEF K B, CHEN W, SHI Y M, et al. The road-map to 6G: AI empowered wireless networks[J]. IEEE communication magazine, 2019, 57(8): 84-90. DOI: 10.1109/MCOM.2019.1900271
- [5] 未来移动通信论坛. 多视角绘6G蓝图 [EB/OL]. (2019-11)[2020-06-28]. <http://www.future-forum.org/dl/191120/08-%E5%A4%9A%E8%A7%86%E8%A7%92%E7%82%B9%E7%BB%986%E8%93%9D%E5%9B%BE.pdf>
- [6] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning: concept and applications [J]. ACM transaction on intelligent systems, 2019, 10(2): 1-19. DOI: 10.1145/3298981
- [7] DU Y Q, YANG S, HUANG K B. High-dimensional stochastic gradient quantization for communication-efficient edge learning [J]. IEEE transactions on signal processing, 2020, 68: 2128-2142. DOI: 10.1109/TSP.2020.2983166
- [8] WANG S Q, TUOR T, SALONIDIS T, et al. Adaptive federated learning in resource constrained edge computing systems [J]. IEEE journal on selected areas in communications, 2019, 37(6): 1205-1221. DOI: 10.1109/JSAC.2019.2904348
- [9] 陈力, 卫国. 未来无线网络下的空中计算技术 [J]. 中兴通讯技术, 2019, 25(1): 29-34. DOI: 10.12142/ZTETJ.201901005
- [10] GASTPAR M. Uncoded transmission is exactly optimal for a simple Gaussian sensor network [J]. IEEE transaction on information theory, 2008, 54(11): 5247-5251. DOI: 10.1109/TIT.2008.929967
- [11] CAO X W, ZHU G X, XU J, et al. Optimal power control for over-the-air computation in fading channels [EB/OL]. (2010-06-17)[2020-06-28]. <https://arxiv.org/pdf/1906.06858.pdf>
- [12] ZHU G X, HUANG K B. MIMO over-the-air computation for high-mobility multi-modal sensing [J]. IEEE Internet of Things journal, 2019, 6(4): 6089-6103. DOI: 10.1109/JIOT.2018.2871070
- [13] CHEN L, ZHAO N, CHEN Y F, et al. Over-the-air computation for IoT networks: computing multiple functions with antenna arrays [J]. IEEE Internet of Things journal, 2018, 5(6): 5296-5306. DOI: 10.1109/JIOT.2018.2843321
- [14] ZHAI X F, CHEN X H, XU J, et al. Hybrid beamforming for massive MIMO over-the-air computation [EB/OL]. (2019-06-08)[2020-06-28]. <https://arxiv.org/abs/2006.04560>
- [15] JEON S W, JUNG B C. Opportunistic function computation for wireless sensor networks [J]. IEEE transactions on wireless communications, 2016, 15(6): 4045-4059. DOI: 10.1109/TWC.2016.2533379
- [16] WU F Z, CHEN L, ZHAO N, et al. Computation over wide-band multi-access channels: achievable rates through sub-function allocation [J]. IEEE transactions on wireless communications, 2019, 18(7): 3713-3725. DOI: 10.1109/TWC.2019.2918145
- [17] RUDER S. An overview of gradient descent optimization algorithm [EB/OL]. [2020-06-28]. <https://arxiv.org/abs/1609.04747>
- [18] YANG K, JIANG T, SHI Y M, et al. Federated learning via over-the-air computation [J]. IEEE transactions on wireless communications, 2020, 19(3): 2022-2035. DOI: 10.1109/TWC.2019.2961673
- [19] ZHU G X, WANG Y, HUANG K B. Broadband analog aggregation for low-latency federated edge learning [J]. IEEE transactions on wireless communications, 2020, 19(1): 491-506. DOI: 10.1109/TWC.2019.2946245
- [20] ZHANG N F, TAO M X. Gradient statistics aware power control for over-the-air federated learning [EB/OL]. (2020-05-08)[2020-06-28]. <https://arxiv.org/abs/2003.02089>
- [21] NARAYANAMURTHY P, VASWANI N, RAMAMOORTHY A. Federated over-the-air subspace learning from incomplete data [EB/OL]. (2020-06-14)[2020-06-28]. <https://arxiv.org/abs/2002.12873>
- [22] KODA Y, YAMAMOTO K, NISHIO T, et al. Differentially private AirComp federated learning with power adaptation harnessing receiver noise [EB/OL]. (2020-04-14)[2020-06-28]. <https://arxiv.org/abs/2004.06337>
- [23] CAI X R, MO X P, CHEN J Y, et al. D2D-enabled data sharing for distributed machine learning at wireless network edge [J]. IEEE wireless communication letters, 2020, (99): 1-1. DOI: 10.1109/LWC.2020.2993837
- [24] WANG F, XU J, WANG X, et al. Joint offloading and computing optimization in wireless powered mobile-edge computing systems [J]. IEEE transactions on wireless communications. 2018, 17(3): 1784-1797. DOI: 10.1109/TWC.2017.2785305

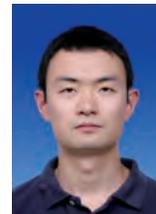
作者简介



曹晓雯, 香港中文大学(深圳)未来智能网络研究院访问学生、广东工业大学在读博士研究生; 主要研究领域为空中计算以及移动边缘计算; 发表期刊和会议论文6篇, 申请专利2项。



莫小鹏, 香港中文大学(深圳)未来智能网络研究院访问学生、广东工业大学在读硕士研究生; 主要研究领域为边缘人工智能以及无人机通信; 发表期刊和会议论文4篇, 申请专利2项。



许杰, 香港中文大学(深圳)理工学院副教授; 主要研究领域为无线能量传输、无人机通信、空中计算和边缘智能等; 先后主持和参加基金项目10余项; 已发表期刊和会议论文百余篇。



万物互联，任重道远

Interconnection of Everything Has a Long Way to Go

李少谦 / Li Shaoqian

(电子科技大学, 中国 成都 611731)
(University of Electronic Science and Technology of China, Chengdu
611731, China)

DOI: 10.12142/ZTETJ.202004008

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20200716.1510.002.html>

网络出版日期: 2020-07-16

收稿日期: 2020-07-11

摘要: 作为 5G 带来的最大的变革, 万物互联面临着很多挑战及不确定性, 需要长时间的探索与开拓, 5G 仅是其探索、起步阶段。在以面向感知与控制为核心的万物互联移动通信中, 5G 将发挥核心系统的作用, 并和其他网络协作发展。先进的网络技术是其发展的重要保障, 而行业信息化的内在驱动力才是其成功的核心要素。移动通信只有与垂直行业深度融合, 共同建立技术链、应用链和生态链, 才能成为行业信息化的支撑技术和基础设施; 而信息通信行业与垂直行业的深度合作与融合取决于新机制、新体制。

关键词: 万物互联; 行业信息化; 垂直行业

Abstract: As the biggest change that 5G brings, interconnection of everything faces many challenges and uncertainties, and it needs to be explored for a long time. As the initial development stage of the interconnection of everything, 5G will play the role of core system and cooperate with other networks in the interconnected mobile communications which focus on perception and control. Advanced network technology is an important guarantee for its development, and industry informatization is the core element of success. Only by deep integration with vertical industries and establishing technology chain, application chain and ecological chain together, can mobile communications become the supporting technology and infrastructure of industry informatization. The deep cooperation and integration between information and communication industry and vertical industries depends on the new mechanism and new system.

Keywords: interconnection of everything; industry informatization; vertical industry

5G 时代已经来临, 在政府的强力政策支持和经济发展需求的驱动下, 中国 5G 初期的建设和发展速度已超过 3G、4G 初期时的速度。面对所取得的成绩, 我们应有清醒的认识: 今天全球 5G 都以提供增强移动宽带 (eMBB) 业务为主, 并会持续很长时间, 但这并不是 5G 的核心目标。5G 带来的最大的变革是移动通信网络从支撑“人与人的连接”为主, 转为支撑“万物互联”, 这是移动通信的革命性变革, 意义重大。5G 万物互联的变革, 面临着巨大挑战, 现在仅处于“孕育期”, 前方长路漫漫。针对超可靠低时延通信 (URLLC) 的 R16 标准才刚刚冻结, 针对海量机器类通信 (mMTC) 的技术与标准难以在短期内确定, 许多技术标准还需到 R17 或者我们称之为 6G

阶段才能完成。5G 独立组网仍存在诸多技术与应用瓶颈, 没有 5G 的独立组网, 5G 就无法提供 URLLC、mMTC 能力。虽然现在面向垂直行业的 5G 万物互联应用, 从政府到各行各业都有着很高的热情和期盼, 各类试点与应用示范层出不穷, 但我们应认识到万物互联面临更多的挑战和不确定性, 需长时间的探索与开拓, 它的发展一定是一个缓慢的、渐进的过程^[1]。

40 年前, 移动通信刚起步时, 我们憧憬着个人通信的理想, 即任何人在任何时间、任何地点与任何一个人实现任何一种媒体的通信; 每个人有唯一的通信号码, 通信的个人性代替通信的终端性。几十年来, 移动通信、卫星通信等技术的飞速发展, 使得人类实现个人通信的理想为期不远了。

人类万物互联的理想看似仅将个人通信理想中的“任何人”改为“任何人与物”, 但这一个字的增加却使理想的实现异常艰巨且更加漫长。5G 仅是万物互联的探索、起步阶段。人类要实现天地万物信息互联的理想, 需要不断地开拓新需求、新技术、新业务、新模式, 建立新的生态链, 需要更多代技术与应用的变革, 任何急功近利都无济于事。

在移动通信走向万物互联的过程中, 通信人需要在以下几个方面努力:

1) 移动通信万物互联行业应用, 将从以人的信息获取与交流为主的消费应用, 转向以信息感知与控制为主的行业应用。新的应用也意味着技术发展的重心将发生变化: 在物联感知与控制的世界里, 互联技术能否用于

精准感知和控制,按需定制、安全可靠、稳健性至关重要,但安全可靠、稳健性常与高速、高效、高复杂度,甚至与高智能冲突;因此,如何实现移动通信万物互联技术的安全可靠、高稳健性是新的挑战。移动通信万物互联的技术发展需将自主可控、高度安全、超高可靠、易定制、易扩展等面向感知与控制的技术作为核心方向。

2) 万物互联的多样性和复杂性远远超过人与人的连接,因此它不可能靠一种通信网络去完成。即便是像移动通信这样的“巨网络”也不行,如同交通运输网络一样,不可能只靠高速公路一种连接方式。5G 将为万物互联提供广域覆盖、可高速移动、大容量大连接的网络支撑,这样的网络将是万物互联的核心系统。核心系统需要与其他系统共同协作,满足不同需求的有线与无线、广域局域与短距离、公众与专用等各种系统的互为补充、互连互通,才会构成万物互联的世界。今天,如果没有 WiFi,即使是 5G 网络也难以承载移动互联网的巨量应用;因此,移动通信与垂直行业都要抛掉 5G 将“一网统天下,实现万物互联”的幻想。5G 万物互联的新架构、新技术可广泛地应用到各类行业专网中。5G 时代,运营商跨界垂直行业,专网和公网将出现更深的融合。移动通信在变革中的目标应为:在万物互联的世界里发挥核心网络作用,与其他网络技术与系统共同发展,共同推进万物互联^[2]。

3) 物联网的发展在中国已持续了 10 余年,政府和各行业投入了大量资源,应用示范遍地开花;但结果并不理想,可谓是“一起一落”。由 5G 掀起的万物互联是物联网发展的又一新高潮,我们应总结之前的经验教训,反思得失,才有可能成功。在物联网 10 余年的发展过程中,成功的应用很多,

失败的应用也不少;但无论成功与失败,却很少取决于网络技术。先进的网络技术仅是万物互联的保障条件,而不是其能否成功的核心要素。

万物互联成功的核心要素是什么?是行业信息化的内在驱动力。行业服务与消费类服务的根本区别是:行业信息服务是行业生产要素中不可分割的核心部分!万物互联不是目的,推动生产效率、产品质量的提升,创造新业务、新形态,扩大新市场,提升行业竞争力才是行业信息化的动能和驱动力。基于信息通信技术的物联网的成功应用都源于内在驱动力。当基于信息通信技术的“有用的物联网”成为了行业中不可分割的核心生产要素,万物互联的理想就离实现不远了。

行业万物互联的内在驱动力需要与行业信息化同步发展,规模化应用取决于行业信息化的发展水平。5G 可为车联网、远程医疗、远程教育、工业互联网、智慧城市等行业规模化应用提供支撑,但这些应用本身尚处在探索起步阶段。任何一个规模化行业要实现万物互联都取决于该行业信息化生态链的进步,越是应用前景广阔的行业,行业互为关联的系统要素就越多,关联就越复杂,建立信息化生态链就越漫长。许多领域的信息化已取得了长足的进步,但离目标仍相当遥远。该如何“破局”?“运动式”的万物互联推动是否还会经历“三起三落”?一切皆有可能!

4) 在这万物互联的新浪潮中,行业信息化是行业生产要素的信息获取、传输、处理、反馈控制与智能生产的一体化,通信技术仅是要素中的一环。移动通信能否成为行业信息化的支撑技术和基础设施,取决于能否与垂直行业深度融合,共同建立技术链、应用链和生态链。行业发生重大进步的信息化过程,需要行业所有生产要素

与体系结构的创新和变革。没有信息通信业与垂直行业的深度合作与融合,创新和变革很难实现。在这变革中,移动通信要从面向公众消费者的“标准设备、流量套餐”式的服务模式,转变到面向垂直行业的“定制化设计、按需服务”的模式。

信息通信行业与垂直行业的深度合作与融合取决于新机制、新体制。没有为垂直行业用户提供真正个性化服务的新生态链,没有与垂直行业共生共荣的新机制新体制,移动通信面向行业的万物互联难以成功。垂直行业万物互联的应用千千万万,不同的行业有不同的需求,难以用一种机制和体制去实现。例如,公共安全、能源、轨道交通等大规模行业,网络安全、管理控制的要求高,需要万物互联自主可控,需要与通信业形成“利益共同体”,联合构建行业可管可控的专业网络。而对分散的、复杂的各种行业应用,公网应创造机制,支持各类万物互联服务企业的兴起,并与专网相结合,为行业提供定制化服务。“八仙过海”推进行业万物互联的应用与发展。

万物互联,任重道远!

参考文献

- [1] 李少谦. 5G: 智能移动通信 1.0 [J]. 中兴通讯技术, 2016, 22(2): 47-48. DOI:10.3969/j.issn.1009-6868.2016.03.010
- [2] 刘翎. 李少谦教授: 5G 既是公网的机会,也是专网的机会 [J]. 专网通信, 2019, 39(5): 53-55

作者简介



李少谦,电子科技大学教授、博士生导师, IEEE Fellow, 通信抗干扰技术国家级重点实验室主任, 国家新一代宽带无线移动通信网重大专项总体组成员, 国家“863”计划 5G 重大项目总体组成员, 国家“973”计划咨询专家组成员, 四川省学术与技术带头人, 国务院政府特殊津贴获得者; 主要研究方向为无线与移动通信技术; 主持完成了 30 余项国家级科研项目, 获国家、国防和省部级科技奖 8 次; 获授权专利 70 余项, 发表论文 200 余篇, 出版专著多部。



网络管理自动化中 闭环形成的概念

Closed Loop in Autonomous Network Management

孟洛明 / MENG Luoming

(北京邮电大学, 中国 北京 100876)
(Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTETJ.202004009

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20200728.1354.004.html>

网络出版日期: 2020-07-28

收稿日期: 2020-06-15

摘要: 通过对网络管理自动化研究发展历程的回顾与思考, 提出了网络管理自动化中闭环形成的概念。闭环形成技术是网络管理自动化研究的基础性工作, 其中有两项关键技术: 网络控制自动化和闭环调用。系统地解决闭环形成问题目前还有很多工作要做。

关键词: 网络管理; 网络管理自动化; 闭环

Abstract: The concept of closed loop in autonomous network management is proposed through the review and reflection of the development course of network management automation research. Closed loop in autonomous network management is the basic work of network management automation research, which includes two key technologies: network control automation and closed loop invocation. There is still much work to do to solve the closed loop problem systematically.

Keywords: network management; autonomous network management; closed loop

降低成本、改善性能和保障运行是网络管理系统建设始终追求的目标。从 21 世纪初开始, 人们逐渐认识到进一步降低成本, 改善性能, 需要提高网络管理的自动化程度。此后, 经历了 20 余年的发展, 相关技术的不断出现持续推动着网络管理自动化技术的进步。

1 网络管理自动化的发展过程

1) 基于策略的网络管理

早期的网络管理自动化是基于策略的网络管理^[1]。该管理方式是指, 在不对网络管理系统重新编码且在其在线运行的情况下, 动态改变网络管理系统, 从而提高网络管理的自动化程度。

2) 基于智能的网络管理

随着研究的不断进展, 人们发现在大规模的网络环境下, 基于策略的网络管理会产生策略冲突, 尤其是在复杂的环境中。网络环境越复杂, 策略冲突就越严重。在这种情况下, 需要提高网络管理自动化的程度。因此, 从 21 世纪初起, 基于智能的网络管理逐渐成为研究热点^[2]。

基于智能的网络管理是指, 通过将网络管理领域专家的经验总结为知识, 形成知识库, 然后基于该知识库进行网络管理。采用这样的方法, 同样可以在不对网络管理系统重新编码且在其在线运行的情况下, 动态改变网络管理系统, 从而提高网络管理的自动化程度。

3) 自主管理的网络管理

随着研究的深入, 人们遇到了和基于策略的网络管理类似的问题: 将网络管理领域不同的专家经验总结为统一的知识表示是一件相当困难的事情。同时, 由于网络管理系统建设具有周期性的特点, 在一个新网络对应的网络管理系统建设的初期, 专家的经验还不能及时被总结出来。

在基于智能的网络管理研究时期, 在欧盟 FP7 和 H2020 的支持下, 研究人员开展了基于自主管理的网络管理的研究^[3]。自主管理的基本思想是让网络本身具有管理能力, 其目标是实现 5S (自感知、自配置、自保护、自优化、自修复), 并在异构无线接入网、软件定义网络 (SDN)、网络

功能虚拟化 (NFV) 等网络上开展具有 2S 或 3S 的初步验证性实验, 同时在 5G 等新型网络上探索具有自主管理能力的体系结构^[4]。有关实验显示: 自主管理能够在自感知方面有较好的效果, 但网络管理自动化程度总体上并无明显提高。

4) 基于深度学习的网络管理

在图像、语音、自然语言处理方面取得重要进展的深度学习方法也逐渐被业界关注。初步的实验表明: 深度学习在故障管理和性能管理预期有比较好的效果, 但总体上网络管理自动化程度并无明显提高, 特别是需要海量的训练数据也是一件比较困难的事情。

以上几种方法的共同特征是:

1) 提出了一种基于 X1 的网络管理自动化方法, 可以很快地将网络管理自动化程度从零提高到一定程度; 但到达一定程度后, 想再进一步提高就显得困难。

2) 又提出了一种基于 X2 的网络管理方法, 又可以很快地将网络管理自动化程度从零提高到一定程度; 但到达一定程度后, 想再进一步提高还是显得非常困难。

以上过程一遍一遍地重复, 似乎存在一个天花板, 只要碰到这个天花板, 自动化程度就很难再提高了。

这个现象引起我们对网络管理自动化的思考: 如果有一个天花板的话,

那么这个天花板是什么?

2 网络管理自动化中闭环形成的概念

在早期, 网络管理面对的是单一网络的简单环境, 执行的是一些简单重复的操作; 而现在, 网络管理面对的是叠加 / 混合 / 综合 / 融合 / 异构的复杂环境, 执行的是一些复杂精细的操作。虽然发生了如此大的变化, 但网络管理过程并没有发生变化, 仍旧是 3 个基本操作: 监视、分析和控制。

实现网络管理过程的一般方法是先对规划的网络管理功能确定管理参数, 再确定管理参数的管理指标, 然后对管理指标进行监视、分析和控制。例如, 在故障管理中, 监视就是故障监视, 采用主动或被动的方式, 实时或周期地收集告警事件; 分析就是故障定位, 根据告警事件进行故障定位; 控制就是故障恢复。如果能够进行故障管理的自动化, 这 3 个操作应当形成闭环, 即故障管理闭环。故障管理闭环的示意如图 1 所示。

目前的各种网络管理方法主要是为了提高这 3 个管理操作中某个操作的质量或效率, 并没有解决管理操作形成闭环这个技术难题。例如, 故障管理的故障监视中采用的各种方法, 就是从海量事件中过滤出告警事件, 然后将大量重复的告警事件收敛为可供分析的有效告警事件。目前使用的

各种方法就是提高过滤和收敛的质量或效率。

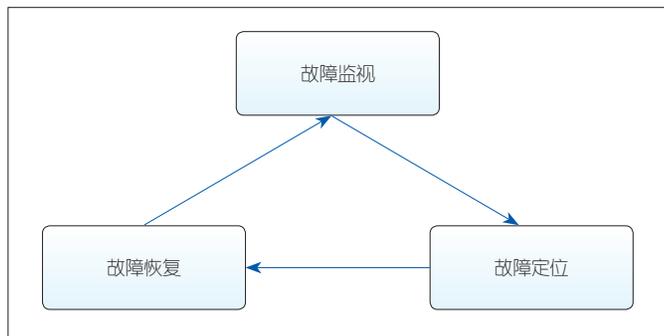
因此, 提高网络管理自动化程度就是要提高形成闭环的程度, 其中有两项关键技术问题: 一是实现网络控制自动化 (网络管理过程中控制操作的自动化), 二是闭环调用技术。

网络控制自动化的难点是网络控制结果存在不确定性。网络在运行的情况下, 特别是在不正常的情况下, 如果改变网络配置, 结果则存在不确定性。目前在自配置、自保护、自修复等方面主要有两种办法: 一种是启动备用 (保护) 部件, 另一种是部件升级。

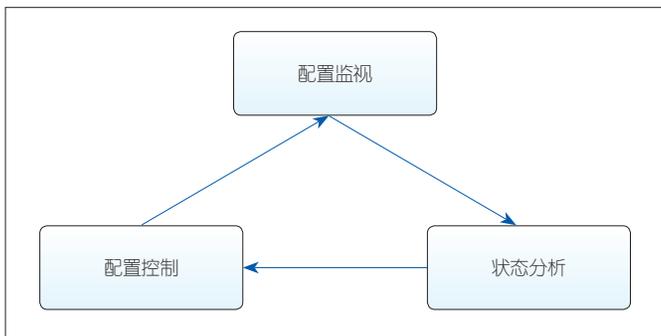
启动备用 (保护) 部件实质上没有对网络配置进行改变, 只是在线更换了相同配置的部件。

部件升级实际上是网络扩容的常用方法。在网络扩容时进行部件升级, 一般选择在网络稳定、低载时有计划地进行。在实施网络控制进行部件升级时, 实现网络稳定和低载, 常用的方法就是部分降级和阻塞部分用户, 但这都有可能产生部件升级原因的正反馈, 从而增加网络控制结果的不确定性。

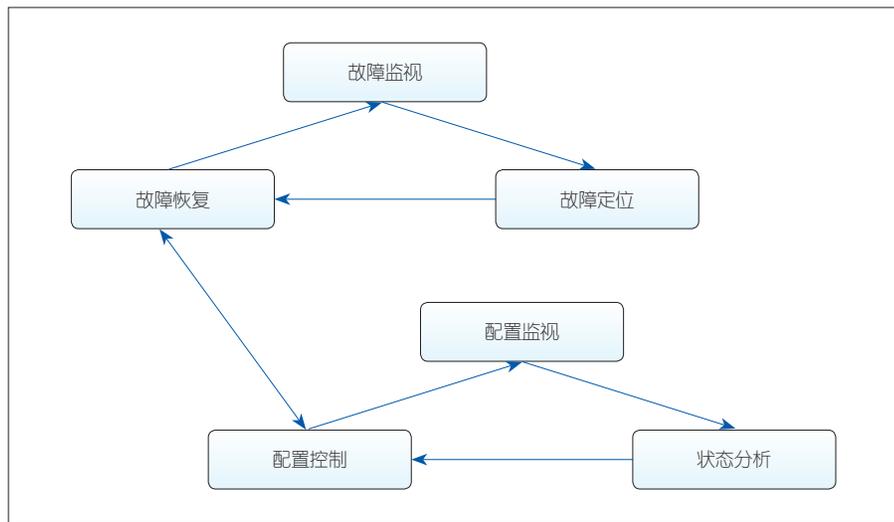
闭环调用是一个闭环调用另一个闭环。例如, 在故障管理闭环的故障恢复操作中, 故障管理的资源中不足以保证故障恢复的质量, 需要通过配置管理改变配置, 以提高故障管理中资源的数量。在这种情况下, 就需要



▲图 1 故障管理闭环



▲图 2 配置管理闭环



▲图 3 闭环调用过程示意图

故障管理闭环中的故障恢复调用配置管理闭环中的配置控制。

配置管理闭环的示意如图 2 所示，闭环调用的过程如图 3 所示。

图 3 显示的是最大管理功能粒度，但实际闭环的管理功能粒度要小得多，因此实际闭环调用也就复杂得多。

通常用闭环调用图来表示闭环调用的关系。在闭环调用图中，结点表示一个闭环，如果一个闭环调用另一个闭环，那么这两个闭环对应的两个结点是邻接结点。

在使用闭环调用图时，闭环调用中的一些判定问题，如闭环调用循环、闭环调用死锁、闭环调用嵌套等，就可以转化为对图或子图的处理。

3 结束语

网络管理自动化研究面临着巨大的技术挑战，迫切需要研究人员开展网络管理自动化方面系统性、长期性的研究，其中基础性的工作是闭环形成技术。系统解决闭环形成问题还有较长的路要走。

参考文献

- [1] COX M D J, DAIVSON R G. Concepts, activities and issues of policy-based communications management [J]. BT technology journal, 1999, 17(3): 155-162
- [2] JENNINGS B, MEER S V D, BALASUBRAMANIAM S, et al. Towards autonomic management of communications networks [J]. IEEE communications magazine, 2007, 45(10):112-121
- [3] EU FP7. EU FP7 project SEMAFOR (Self-management for unified heterogeneous radio access networks)[EB/OL].[2020-07-07]. <http://fp7-semafour.eu>
- [4] NEVESA P, CALEA R, COSTAA M, et al. Future mode of operations for 5G - The Selfnet approach enabled by SDN/NFV [J]. Computer standards & interfaces, 2017, 54(4): 229-246. DOI: 10.1016/j.csi.2016.12.008

作者简介



孟洛明，北京邮电大学教授、博士生导师，北京邮电大学学术委员会主任，国家级有突出贡献的中青年专家，国家杰出青年科学基金资助获得者，“长江学者计划”特聘教授，国家“973”计划项目首席科学家，国家自然科学基金创新

研究群体带头人；长期从事通信网和网络管理方面的研究工作；研究成果获国家科技进步二等奖 2 次。



通信产业发展回顾与展望

Review and Prospect of Communications Industry Development

张云勇 /ZHANG Yunyong

(中国联通集团产品中心, 中国 北京 100032)
(China Unicom Group Product Center, Beijing 100032, China)

DOI: 10.12142/ZTETJ.202004010

网络出版地址: <https://kns.cnki.net/KCMS/detail/34.1228.TN.20200727.1730.002.html>

网络出版日期: 2020-07-28

收稿日期: 2020-06-29

摘要: 从通信产业的发展与演变入手, 阐述 5G 网络架构演进、天地一体化通信网络以及多样化、定制化通信终端。未来面向传统行业和新兴行业的各类创新应用场景, 通信产业发展空间将被不断扩宽。在共建、共享、合作的趋势下, 产业各界共同打造更稳健、更安全、更有弹性、更智能的通信网络。在可持续健康发展的同时, 通信产业进一步推动社会经济高质量发展。

关键词: 通信技术; 5G 网络; 终端; 应用; 发展机遇与挑战; 高质量发展

Abstract: Starting from the development and evolution of the communications industry, the evolution of 5G network architecture, the space-ground integrated communication network, and the diversity and customization of communication devices are discussed. In the future, the communications industry will continue to accelerate the evolution to fulfil the requirements of various innovative application scenarios from traditional and emerging industries. In the trend of co-construction, sharing, and cooperation, various sectors of the industry are building the communication network with a more stable, safer, more flexible, and smarter infrastructure. The sustainable and healthy development of the communications industry is essential while promoting the high-quality development of economy.

Keywords: communication technology; 5G network; terminal; application; development opportunities and challenges; high-quality development

1 通信产业发展回顾

1.1 技术更迭拓展通信产业深度和广度

1G 实现了移动通话, 2G 实现了短信、数字语音和手机上网, 3G 带来了图文并茂的移动互联网, 4G 推动了移动视频的发展, 5G 支持虚拟现实 (VR)、增强现实 (AR) 及高清视频传输, 有望实现海量物联。从 1G 到 5G, 通信技术更新换代越来越快, 技术的快速发展有效地促进了通信产业的多元化发展, 为通信设备商、芯片商、通信运营商、软件制造商、终端厂商等产业上下游带来了良好的发展机遇。

经过多年发展, 中国通信产业服务已经发生根本性的变化——从单一

化的通信服务提升到综合化的信息服务。通信行业已经成为当今基础民生服务行业之一, 并渗透到人们生产生活的方方面面。通信产业与传统产业的快速融合, 拓展了通信产业的发展空间, 同时也为农业、工业、服务业等传统行业的可持续发展提供新机遇、新空间。

1.2 网络扁平化、云网一体化带动 5G 网络架构变革

在通信网络从 2G 逐步演进到 5G 的过程中, 原来层次化的通信网络架构也变得越来越扁平化。扁平化网络架构有利于构建低时延、低成本的通信网络, 有利于提升 5G 通信网络的

稳定性、网络容量、服务效率与质量, 有利于实现 5G 高带宽、低时延、高可靠和海量物联等应用场景。未来随着人工智能、大数据、区块链等新型信息技术的引入, 通信网络架构将会更加扁平化, 通信网络的网内和网间的协作效率都将会得到进一步提升^[1]。

伴随云计算技术的进步, 以及在网络功能虚拟化 (NFV)、软件定义网络 (SDN) 和人工智能等技术和基础设施的共同驱动下, 云网一体化正在成为趋势。云网一体化融合云计算、通信、IT、大数据、人工智能、区块链等诸多新技术和新产业, 具有智能化、自服务、高速、灵活等优势, 可以为通信网络带来新一轮架构变革。

通信运营商正在研究和尝试云网一体化建设^[2]，以期依托云网一体化服务来推动自身网络资源的优化升级，实现网络与云的敏捷协同、按需互联，提升自身的通信网络质量和通信业务能力。

1.3 天地一体化通信网络逐步形成

从电缆到光缆、从有线到无线、从模拟到数字，近年来中国通信网络建设综合实力持续增强，已经在光网建设、移动网络建设、大数据基础设施建设等方面取得了巨大成就。中国通信网络规模和质量，以及通信服务能力都走在世界前列。

中国幅员辽阔，但受制于自然环境因素，还有不少地区难以通过固定通信网络或移动通信网络提供有效的网络覆盖和通信服务。通信运营商仍在不断扩建和完善海底光缆、陆地光缆和移动通信网络，努力增加网络覆盖能力和网络服务质量，同时也在积极推进光网络、移动通信网络、卫星通信网络和北斗卫星导航系统的深度融合，以提供天地一体化的通信网络服务和精准定位服务。

卫星通信网以其日益凸显的国家战略地位、潜在的市场经济价值、稀缺的轨道频谱资源，正在成为各国战略布局和竞争的焦点。利用“低轨卫星+5G”构建天地一体化信息网络，正在成为各国科技竞争的新战场。美国太空探索公司（SpaceX）从2015年开始规划“星链（Starlink）”计划，预计发射上万颗卫星以组成庞大的空中通信网络，目前成功发送了数百颗小型通信卫星，已经具备了初步的卫星通信网组网能力。欧洲与中国卫星通信相关公司也陆续推出了自己的卫星通信网计划，在低轨卫星通信领域有所尝试。2020年初，卫星通信网已纳入中国“新基建”范畴，成为中国

通信网络基础设施的重要组成部分，这将为我国卫星通信网的建设和发展带来新的机遇和动力。

可以预见，未来的5G通信网络将支持与各种地面无线移动通信网络、中高低轨道的卫星移动通信网络以及短距离无线通信网络之间互联互通和相互协作，海陆空一体化的通信网络 and 全空域的通信服务将逐步成为现实。

1.4 终端多样化、个性化应用于更多场景

相对于20世纪的电话、BP机、普通手机等功能单调的通信终端，随着通信网络的快速更新迭代，通信终端不再仅仅局限于通信功能，开始与日益丰富多彩的业务相融合，变得越来越多样化和个性化，出现了大量类似于智能手表、智能家居、车载终端等非手机形态的移动通信终端。据统计，当前非手机形态的移动通信终端新产品占据中国终端新产品市场的6成左右，且比例仍在持续增加，主要集中在可穿戴设备、智能车载终端、工业互联网设备终端等领域。在可穿戴设备方面，5G网络为语音交互、视频对话、在线音乐等应用提供了坚实的数据传输基础，丰富了可穿戴设备的产品种类；在车载终端方面，自动驾驶和车联网等业务需求推动了车载终端进一步迈向智能化、个性化和多样化；在工业互联网方面，5G海量互联和高可靠的宽带数据传输能力有利于部署和应用更多具备自感知、自学习、自适应、自控制能力的工业终端^[3]。

同时，在零售、餐饮、物流、仓储、金融、教育、医疗等领域也涌现了大量新型通信终端。随着人工智能、区块链、物联网、嵌入式用户识别卡（eSIM）等技术的发展和成熟，通信终端将会更多地融入到我们的生产和生活之中，多样化和个性化也将是通

信终端的发展趋势。

1.5 跨界融合应用拓宽通信产业发展空间

过去通信产业应用单一，主要功能是拨打电话、收发短信等。随着移动互联网的发展，通信产业应用跨界融合速度不断加快，飞速发展的通信网络为更多行业提供了新的可能。5G大带宽、低时延、广连接的3大业务场景，在城市管理、民生服务、智能制造等方面均有大量应用。在5G时代，信息化与工业化将会走向深度融合。5G为工业互联网提供端到端毫秒级时延和接近100%的高可靠性通信保障，可以满足95%以上的工业大数据无线传输和实时处理需求，有利于推动传统工业向着数字化、网络化、高效化、智能化方向发展。

5G灵活、开放、高效融合的特性使得5G创新业务在工业、农业、交通、教育等各个行业遍地开花，助力垂直行业应用更加丰富多彩。在未来通信时代，通信技术将继续以用户体验为导向，满足不同终端、不同用户个性化和定制化的应用需求。通信运营商和传统企业也将会持续在商业模式上进行创新，更加积极、主动、及时、智能地满足多样化的行业需求。此外，5G网络也将持续和大数据、人工智能、云计算、区块链、物联网、工业互联网等技术和基础设施相结合、相互赋能，形成更加多样化的5G融合应用，推动通信产业的可持续发展。

1.6 共建、共享促进通信产业可持续发展

5G给通信产业带来巨大的发展机遇，但发展5G也会面临极大的挑战。首先，中国幅员辽阔，人口众多，全面建成5G网络所需的基站数量多、投资量大。据测算，未来5G基站数量

将是4G基站数量的2~4倍,这意味着通信运营商在5G基站的建设工作中需要投入更多的资金。在4G成本尚未完全收回的情况下,建设5G网络将不可避免地给通信运营商带来巨大的资金压力。其次,移动通信频谱资源的稀缺性也制约着5G的未来发展。全球通信网络可用的频谱资源十分有限,未来通信网络将面临网络容量不足、频谱效率低、网络覆盖不高等一系列问题。这些问题的解决除了通过腾退2G和3G频谱、开发新5G频段外,还需要通信运营商在提升频谱利用效率上共同努力。再者,5G的高能耗也是通信运营商不可忽视的问题。据测算,一个5G基站的功耗是4G的2~4倍,之前提到的5G基站的数量同样是4G的2~4倍。这意味着在相同的覆盖范围下,极端情况下5G的功耗和维护成本将会是4G的10多倍,这对于通信运营商而言将会是非常巨大的成本开支。最后,在行业竞争方面,对整个通信产业而言,全球通信产业竞争压力正在持续增大,通信市场日益饱和,通信运营商所面对的行业内外的竞争将更加激烈。

面对目前通信产业快速发展带来的机遇与挑战,全球通信运营商都在采取积极的应对策略,包括跨通信运营商的深度共建共享、跨网协作等,以期促进通信产业的可持续发展。具体来说,在提升通信频谱利用率方面,未来通信运营商将会采取动态调配频率资源,多家通信运营商、多种制式通信网络共享频率资源等方式来提升频谱利用率。此外,通信运营商也正在研究通过引入区块链、人工智能、大数据等新型信息与通信技术(ICT)和基础设施,助力未来网络的建设、运营、维护和管理,从而进一步促进通信网络的共建和共享。在网络建设方面,面对更高的建网投资、更快的

建网需求,通信运营商的共建共享需要持续深化。通信运营商正在通过5G共建共享来降低5G网络建设和运维成本,提高5G网络覆盖率,以期快速形成5G服务能力,提升5G网络效益和资产运营效率,达成互利共赢。中国联通与中国电信已经达成合作协议,正在共同建设5G网络,共建共享一张5G接入网,旨在通过深度合作达到5G网络建设和运营的可持续发展。中国移动和中国广电近期也签署了5G共建共享合作框架协议^[4],开展5G共建共享以及内容和平台的合作。可以预见,在资源和成本的双重压力下,通信运营商之间的5G通信网络共建共享也将成为全球通信网络建设的大趋势。

2 通信产业发展展望

通信技术助力社会从信息化时代进入智能化时代。从1G到5G,通信技术和通信服务更新换代越来越快,网络带宽越来越大,传输速度不断提升,应用场景逐渐丰富。目前,通信技术已经基本解决人与人之间的通信需求,正在朝着解决物与物之间通信需求的方向发展。5G技术的商用进一步推动了经济社会发展和产业融合创新。5G和物联网、人工智能、边缘计算、区块链等新技术相互融合、协同发展,将持续提升数字化智能化水平,为社会生产和生活注入新的活力。

面对机遇与挑战,通信产业需要探索5G时代全新的发展模式,要在助力5G生态繁荣发展的同时实现自我生长。在网络发展方面,通信产业将持续展开深度共享共建,实现跨网协作,达成互利共赢,共同促进5G网络可持续发展;在行业发展方面,通信产业也将持续与产业链上下游展开密切合作,拓展行业渗透领域,在工业、农业、商业等多领域发掘自身潜力,

促进整个通信产业的可持续发展^[5]。

3 结束语

随着通信技术的不断提升,未来的通信网络将更稳健、更安全、更有弹性,网络容量、服务效率与质量将更能满足各行各业信息化和现代化发展的迫切需求。通信产业将实现海、陆、空一体化的移动通信网络,实现全球泛在覆盖的高速宽带通信,实现万物互联。作为信息化建设基础产业,中国通信产业将持续为中国和全球数字经济的发展提供强劲动力,为社会、经济高质量可持续增长和发展提供坚实基础。

参考文献:

- [1] 王翠林. SA网络架构是5G商用最终目标[N]. 通信产业报, 2020-06-01(7). DOI: 10.28806/n.cnki.ntxcy.2020.000183
- [2] 王小雨, 贾宝军, 徐雷. 云网一体赋能运营商数字化转型[J]. 信息通信技术, 2019, 13(2): 20-25. DOI: 10.3969/j.issn.1674-1285.2019.02.004
- [3] 少宇, 王凡, 曹方. 迎接新基建浪潮, 推动5G产业高质量发展[J]. 网络安全和信息化, 2020, (06): 32-35
- [4] 中国移动与中国广电开展5G共建共享合作[J]. 电信工程技术与标准化, 2020, 33(6): 48
- [5] 王雪梅. 5G产业正踏实前行[N]. 人民邮电, 2020-06-16(5). DOI: 10.28659/n.cnki.nrmtyd.2020.001461

作者简介



张云勇, 中国联通集团产品中心总经理、原中国联通集团研究院院长, 国务院特殊津贴专家, 百万人才工程国家级人选, 第十三届全国政协委员, 金砖五国工商理事会数字经济与放管管制组成员, 北京邮电大学、北京交通大学兼职教授, 中国通信学会、中国电子学会、中国计算机学会等多个学会会士, 《通信学报》《电信科学》等期刊编委, 央企安全联盟智库首批专家, 获得中国有突出贡献中青年专家称号; 自主研发沃云平台、海量上流量查询分析系统, 制定了国际上首个云计算框架以及网络即服务标准; 获工信部ITU优秀文稿奖2次、优秀个人奖2次、中国通信学会科技进步奖25次。



知识 + 数据驱动学习： 未来网络智能的基础

Knowledge-and-Data Driven Learning: Foundation of Future Network Intelligence

朱近康 / ZHU Jinkang

(中国科学技术大学, 中国 合肥 230027)
(University of Science and Technology of China, Hefei 230027, China)

DOI: 10.12142/ZTETJ.202004011

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20200727.1732.004.html>

网络出版日期: 2020-07-28

收稿日期: 2020-07-01

摘要: 讨论了未来网络智能的核心问题和学习能力, 提议采用知识 + 数据驱动学习模型作为未来网络智能的基础。论述了学习模型中的知识驱动和数据驱动及反向传播判决, 以及利用该模型实现未来网络的智能管控方法。

关键词: 未来智能网络; 知识 + 数据驱动学习模型; 智能管控

Abstract: With the discussion of the central issues and the learning capabilities of future network intelligence, it is proposed to use the knowledge-and-data driven learning model as the basis of future network intelligence. The knowledge-driven learning, data-driven learning and back-propagation decision in the learning model are introduced, and the intelligent management and control methods for future networks are given.

Keywords: future intelligent network; knowledge-and-data driven learning model; intelligent management and control

1 未来网络面临的挑战

对未来无线网络 (5G 和 6G) 所面临的挑战, 已经有较多研究者从不同角度进行了研究。本文中, 我们就未来无线网络智能管控所面临的重大技术课题进行讨论, 研究值得特别关注的核心和基础, 提出知识 + 数据驱动学习模型作为解决之道。

1.1 未来网络智能 3 大核心

未来网络 (5G 和 6G) 是在现有网络的基础上发展起来的。它既要包容既有网络和技术, 比如 4G、5G、物联网 (IoT) 及各种专用网, 又要引入和构建新的 6G 网络, 这使未来网

络体系变得极为繁杂。此外, 各种新型业务和服务层出不穷, 与普遍应用的软件定义网络交叉叠加, 使得人们不得不寻求未来网络管控的技术基础和实现方法。由于需要诉求、网络变量和资源开销都比较多, 如果没有新技术基础和新技术方法, 这个课题的发展会相当艰难。

因此, 未来智能网络必须具有极其简单的网络架构, 把过去、现在和未来的可能手段都统合起来, 充分利用智能学习方法实现各种资源的最优化利用。

未来网络智能将会有 3 大核心: 智能管控的极其简洁的网络架构、知识 + 数据双驱动的学习机制、全场景

全业务动态联合优化。其中, “管控”是目标, “学习”是基础, “优化”是方向, 这 3 点在未来网络智能中缺一不可。

1.2 未来网络智能的 3 大能力

1) 全场景管控。面对地面覆盖、空中覆盖、天体覆盖以及微覆盖和点覆盖的全场景应用, 未来网络的管控能力要能够有效实施、综合利用。现在地面网络 5G、4G、IoT、WiFi 和个人网等统合起来的超密集网络, 要做到优化管控已经十分艰难。在未来, 如果没有强大稳健的学习能力, 地面与空中切换、地面与卫星交互是无法智能管控的。

2) 全知识学习。作为未来网络智能管控的基础手段,全知识学习会利用直接和间接的各种知识,在大数据配合下自主、独立、透明地学习。全知识包括用户传输需求、用户属性、网络参数、资源开销等。用户传输需求可能比较简单,但伴随的用户属性可能相当多样。在未来,这些知识的利用可能是实现期望目标的重要因素。

3) 全透明优化。它涉及学习的透明、接口的透明和运行的透明,其中学习的透明尤其重要。打开人工智能(AI)的“黑匣子”,使现有的深度学习和AI变成可解释的学习模型,是我们必须面对和解决的问题。做到了学习透明,接口透明和运行透明就有了实现基础。

1.3 未来网络的智能学习

智能学习,通常是指机器学习、深度学习、强化学习,甚至是这些学习的结合,它是数据驱动的学习类别。在输入对象和输出对象一致的情况下(如新媒体学习),这类学习大多十分有效,能够提高辨识能力。这类学习可被称为信号处理型学习。

2017年一篇《模型驱动的深度学习的深度》论文,开启了模型+数据驱动学习模型的研究和应用。2018年用于多输入多输出(MIMO)检测的模型驱动深度学习网络、2019年用于波分多址的模型驱动深度学习的方法以及后用于物理层传输的模型驱动深度学习的方法等,都是该领域典型的进展。由于输入与输出的对象基本上相同,这些研究进展可以归并为信号处理类型学习,并且这些进展的研究重点主要在提高判决能力上。

对于未来网络的智能管控,我们要把各种知识和参数的相互关系和联动看成一个学习整体。通常,输入的是用户需求和用户属性,输出的是优

化的网络参数和必须提供的资源开销。随着用户需求和用户属性的增多,智能管控变成一个极其复杂的优化问题。这就要求智能学习要有复杂群体的设计能力。为此,我们提出了一种开放、透明、可解释的深度学习方法,即知识+数据驱动的学习方法^[1],这种方法可以把知识驱动与数据驱动有机结合起来,能够融合无线知识的有效性和数据的动态实时性。

2 知识 + 数据驱动学习模型

知识+数据驱动学习模型,集中体现了无线通信网络智能化的知识+数据双驱动的学习架构、知识和数据分别学习和训练的数学表达、统一的学习输出和损失率反向传播判决。

知识+数据驱动的学习模型,首先涉及的是知识的范围和表达,随后是知识驱动和数据驱动在各层的分与合以及它们透明的可解释性,最后是反向传播损失率判决和反馈实施。学习模型把这些综合起来,就形成了一种依据用户业务需求、充分利用用户属性和无线大数据、自动学习知识变量取值和使用最小资源开销、实现最佳通信的能力。

2.1 全知识的学习范畴和表征

无线通信和网络的全知识分为3个范畴:用户面知识、网络面知识、服务面知识。每个面有各自的知识变量和公式表达。

1) 用户面知识。用户面知识最直接的是用户和用户传输需求。单位面积用户数高达上千万、带宽从吉比特每秒变为到太比特每秒将是常态。这是从1G到5G一直面临的要求,到5G和6G,甚至以后,也不会例外。用户属性是用户面知识的一个重要方面,也是未来必须引入和应用的。用户属性具有知识性,相互关联而又随

机不同的^[2]。随着通信从服务“人”走向服务“物”、再走向服务“虚拟对象”,针对用户和用户传输需求,需要寻找它们附带的不同用户属性,以便精准估计用户意图,实现准确送达。当然,除用户和用户传输需求之外,用户的属性还包括用户移动速度、用户数据时延要求、用户所在位置等。这些用户属性都可以用数学量来表达,以实现在智能网络中得到最大限度的利用和满足。

2) 网络面知识。网络面知识包括无线传输、无线接入、网络配置、定点送达等。网络面知识涉及的网络类型问题有:是4G、5G,还是6G?是IoT、WiFi,还是picocell?网络面知识还涉及传输与网络参数问题:是正交准正交传输,还是干扰对消或对齐传输?是大区无缝覆盖,还是密集热点覆盖?是MIMO天线,还是超表面多天?这些知识可用数学符号来表达。有些已经被证明有效可靠的数学公式,可作为智能构建复杂网络和提高管控能力的知识对象^[3]。

3) 服务面知识。服务面知识主要是指完成给定服务所需的资源开销,如频谱频段和带宽、峰值功率和平均功率、算力开销和缓存大小、人力开销和经济成本等。这些知识是实现未来网络智能优化管控所必须承受的开销,我们应尽量使之最小化。

2.2 知识驱动 + 数据驱动的双重结构

知识+数据驱动学习模型,由数据驱动的深度学习架构和知识驱动的逐层优化过程组成,按数据驱动和知识驱动同时推进和彼此交互来完成。

数据驱动的深度学习,同标准深度学习完全一样,有输入层、隐含层、输出层。层间通过学习加权函数连接,不同层的节点有不同层运算。数据驱动的深度学习最后把输出层给出的损

失率作为反向传播判决变量，做反馈深度学习。数据驱动的深度学习输入是用户传输需求，可以是用作分析和预测的历史大数据，也可以是用作配置和优化的动态实时用户大数据。

知识驱动的逐层优化，是在数据驱动深度学习的伴随和支持下进行的。不同层使用不同的知识变量，这是依据通信网络知识来确定的。知识变量的取值，可从数据驱动学习过程中获得。当然，对于非随机变动不可准确预测的情况，知识变量的取值也可以从已经被证明行之有效的公式和运算中获得。

因此，知识 + 数据驱动学习模型，构成了知识驱动 + 数据驱动的双重体系结构，可以实现知识学习和数据学习的交互运算、融和运行。

2.3 知识参与的深度学习

知识 + 数据驱动学习模型的每一层，都有数据输入和知识输入。数据输入输入的是用户、用户传输需求，以及用户的不同属性。其数据值是随机、不可当即准确预测的。知识输入输入的是选取的知识变量和数学表达。各层知识变量的选择，按用户属性、网络参数、资源开销分层布局。知识变量的设置因要解决的问题不同而有所区别，但知识变量的取值，可从数据学习或数据支持下的运算中获得。通常，这会有一些约束和规范。

如前所述，在深度学习的前向结构中，各层按数据输入和知识输入双重推进。首先，在数据输入时，通过对加权矩阵元素值的调整获得的可信输入加权进入各层节点。在前向学习中，知识变量通过输入数据计算的关联公式获得取值。通过输入数据对输入加权矩阵和知识变量取值的学习，完成由输入数据、加权矩阵和知识变量参与的节点运算，随后输出本层数

据和知识变量。最后，在输出层输出学习结果和反向传播判决损失率，支持反馈学习和输出。

深度学习的反向传播，是指利用输出层输出的损失率，相对于数据加权矩阵和知识变量作偏微分，获得加权矩阵和知识变量最佳反馈调整途径，然后再一次做前向学习。据此，综合前后各层整体的不断学习，反复进行，直至达到期望目标，实现复杂网络最佳管控。

2.4 知识 + 数据驱动学习模型是可解释的

在知识 + 数据驱动学习模型中，各层的输入是从用户传输需求数据开始的。前一层的输出数据送到下一层加权处理后，被作为节点输入数据，进行逐层计算，直至输出。各层的知识变量是根据各层学习功能和优化需要来明确的，同时各层彼此不同；但知识变量的取值，可通过对输入数据的有规则学习来确定。

由此可见，知识 + 数据驱动学习模型的各层功能是清晰的，学习过程是明确的。数据学习的走向受到知识变量的约束，知识变量的取值由输入数据演变而定。每一层在推进中将不会存在人为干预或人工操控的可能。因此，知识 + 数据驱动学习模型是可解释、透明的深度学习模型。

3 未来网络的智能管控

3.1 未来智能网络的极简架构

未来智能网络是一种智能化的极简网络，也是一种分层学习架构。它包含一个由终端网络、接入网络、核心网络构成的三层基础架构，拥有明确的网间输入输出接口、终端用户需求数据输入和核心网络服务输出应用。这种分层学习架构，可以被定义为知

识 + 数据驱动学习的虚拟体系架构。网间接口涉及层间输入、输出和加权矩阵的设计，以完成网间交互。面对用户需求和用户属性，通过学习和聚类，送往上层网络，或者驻留本层网络，来完成端到端服务。

每层网络又是相对独立的，可按各层网络功能构成自己的分层或分块结构，并等效为另一类知识 + 数据驱动学习模型。根据用户提出的传输需求和用户属性，各层网络可以自行优化设计，实现本层网络连通的低时延和低功耗的端对端服务。

3.2 智能管控的动态联合优化

面对各种动态随机的用户需求，未来网络的智能管控能够实时动态调整参与服务的网络架构和参数，实现各层网络资源的开销最小、效益最大。未来网络的智能管控能涉及的关键词是动态、联合、优化。

未来网络的智能化是实现网络整体管控的动态联合优化。知识 + 数据驱动学习模型和运行规范，是未来网络智能的基础。用知识 + 数据驱动学习模型作为基础模块，在相对长时间的动态承载活动帧内，可构建一套网络整体管控软件。动态承载活动帧包括智能网络的初始化、小于网络能力的欠载、大于网络能力的过载和随机需求动态联合优化。

动态承载活动帧的长短，不仅取决于网络面对的大量用户需求的统计特性稳定期，还取决于智能网络最基本变量的持续期。无线网络、频谱带宽和小区半径，都有较长的稳定期。快，可按小时计，如终端网络；慢，则按天或月计，如接入网络，有的甚至按年计，如核心网。

这里以无线网络为例，来说明未来网络智能管控的动态优化。首先要实现无线网络智能管控的初始化，即

利用过去的用户需求大数据和它们的相关属性，以及网络最基本变量（如频段和小区半径），做初始化训练。获得的动态承载活动帧的基础值，在动态承载帧内不容易变动。

对小于网络能力的欠载情况，在初始化确定的无线网络最基本变量基础上，本次承载给定的频谱带宽有富余。这时，有的主动承载上一次未能完全传输的需求，或减少网络的密集程度、降低发送功率、回退 MIMO 天线数目，以使网络的功率开销最小。

对大于网络能力的过载情况，当无线网络的所有资源效率和开销放到最大但仍不能满足到达的海量传输需求时，不得不按用户传输需求的时延分类，把可承受较长时延的需求从此次传输中转移给下一次或再下一次的传输。这样的联合协作和最大化频谱利用能力，可有助于实现用户需求传输最大化。

随机需求的动态联合优化，就是

把初始化、欠载传输功耗最低、过载传输频谱效率最大整体结合起来。无线大数据的支持，使网络基本变量得选择变得更精准，也使后续实时动态学习调整的代价变得最小。在动态承载帧内，欠载和过载会交替出现，并动态协作。二者联合起来能够做到传输能力最大、所需资源最少，有助于实现智能网络管控目标。

4 结束语

综上所述，未来智能网络可以构建成为一个适合深度学习的极简网络架构。基于知识 + 数据驱动的学习模型，能够实现网际间的学习优化和网络内的学习优化，并构成一个动态联合优化体系，有助于实现未来网络的智能自主管控。

参考文献

- [1] ZHU J K, ZHAO M, ZHANG S H, et al. Exploring the road to 6G: ABC—foundation for intelligent

mobile networks [J]. China communications, 2020, 17(6): 51–67

- [2] ZHU J K, GONG C, ZHANG S H, et al. Foundation study on wireless big data: concept, mining, learning and practices[J]. China communications, 2018, 15(12): 1–15

- [3] ZHU J K, ZHAO M, ZHOU S L. An optimization design of ultra dense networks balancing mobility and densification [J]. IEEE access, 2018, 6(1): 32339–32348. DOI: 10.1109/ACCESS.2018.2845690

作者简介



朱近康，中国科学技术大学教授、博士生导师，曾任中国科学技术大学信息科学技术学院常务副院长、校学术委员会副主任，国家科技部“863”计划通信主题专家组成员、个人通信专家组组长，亚太地区移动通信技术论坛中国

代表，2009年“Green Wireless Technology and System”黄山学术会议执行主席，2014年起每年担任“无线大数据研讨会”的执行主席；长期从事无线移动通信技术和系统的研发工作，近年来主要从事绿色无线通信技术和网络、无线大数据和无线 AI、无线通信基础理论等方面的研究。



针对 5G/B5G 的大规模 MIMO 系统射频前端设计

RF Front-End Designs of MIMO Systems for 5G and Beyond

马建国 / MA Jianguo

(广东工业大学, 中国 广州 510006)
(Guangdong University of Technology, Guangzhou 510006, China)

DOI: 10.12142/ZTETJ.202004012

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20200804.1826.002.html>

网络出版日期: 2020-08-05

收稿日期: 2020-07-08

摘要: 大规模多输入多输出 (MIMO) 成为支撑 5G/B5G 同时满足大数据率和大空间覆盖的必选技术, 包含天线阵和射频驱动的射频前端设计技术成为 5G/B5G 时代的关键技术之一。成本效益将主导大规模 MIMO 系统的商业化, 面向成本的设计 (DfC) 是新的趋势。介绍了当前针对射频前端系统 DfC 的最新进展和引入大规模 MIMO 后带来的新挑战, 指出如何修正传统的香农定理成为 5G/B5G 时代的最基础理论挑战。

关键词: 大规模多输入多输出; 射频前端设计; 面向成本的设计; 5G/B5G; 香农定理

Abstract: Massive multiple input multiple output (MIMO) has become the indispensable technology for 5G/B5G in order to support high-speed data rate with broader spatial coverage concurrently, and the design methodology of radio frequency (RF) front-end including antenna-arrays and RF driver is becoming one of the key technologies. Costs efficiency will dominate the commercialization of massive MIMO systems, and design for cost (DfC) is the new trend. Recent progress of DfC for RF front-end system based on DfC and new challenges brought by the introduction of massive MIMO are discussed. It is pointed out that how to modify the traditional Shannon Theorem becomes the most basic theoretical challenge in the 5G and beyond era.

Keywords: massive MIMO; RF front-end design; design for cost; 5G and beyond; Shannon Theorem

1948 年, 香农发表著名文章《A Mathematical Theory of Communication》, 奠定了现代信息论和现代通信理论的基础。他提出了著名的香农定理^[1]:

$$C = B \times \log_2(1+P/N), \quad (1)$$

其中, C 是对于频率带宽为 B 的通信通道所能够支撑的最大数据率, P 是信号的功率, N 是该通道的噪底。公式 (1) 告诉我们: 只有不断提高信号功率和信道的频率带宽, 才能获得更高的通信速率。香农定理当初是针对点对点的有线通信系统提出来的。如果不考虑线路的损耗, 那么接收到的功率就是发射的功率。进入无线和移

动数字通信时代, 公式 (1) 依然是当前业界用来计算无线通信系统最大信道容量的最基础公式, 但对于无线和移动通信来说, 接收端天线所接收到的功率仅仅是发射端天线辐射出来功率的一小部分。通常假设接收天线都是点源并且相互处于远区场, 令 P_T 为发射天线辐射出来的总功率, P_R 是接收天线接收到的功率, G_T 和 G_R 分别是接发收天线的增益, L 是无线电波在接发收端之间总的空间等效路径衰减, 则有:

$$P_R = P_T G_T G_R L. \quad (2)$$

将公式 (2) 代入公式 (1), 可

得到修正后的针对无线和移动通信的最大信道容量:

$$C_w = B \times \log_2(1+P_R/N). \quad (3)$$

由公式 (3) 可知, 要进一步提高无线通信速率只有两种办法: 加大信道带宽或者提高接收端天线接收到的功率。由于受物理规律限制, 同时频谱资源又极其受限, 空间的路径衰减无法减少, 发射端的发射功率又不能够无限制地提高。从技术上来说, 只有提高接发收天线的增益才是最切实可行的途径。方向性天线可以提供一定的天线增益, 但高增益天线的空间覆盖则会相应地急剧减小。只有增

加天线数量，才能够同时满足 5G/B5G 在高速率和大空间覆盖的双重需求。可以同时实现多波束赋形和空间波束扫描的大规模多输入多输出（MIMO）将成为支撑未来 5G 和 B5G 的最核心关键技术之一。

当前最流行的大规模 MIMO 包含 $N \times M$ 个平面天线单元的有源天线阵，每一个天线单元由一个独立的射频链路来驱动。由于实际的应用环境千差万别，平面 MIMO 不仅在空间分布上存在波束死角，而且由于终端平台本身也存在动态变化。这使得波束无法始终瞄准预设的方向，比如船舶上的卫星天线阵的波束必须始终指向卫星，但船的波动和倾斜使得平面 MIMO 系统无法应对很多波动状态。特别是对于舰船和航空器等移动平台，三维随机动态摆动将使平面 MIMO 阵列很难确保波束保持在所希望的空间指向上。由于三维 MIMO 可以解决这些难题，所以三维 MIMO 阵列也就成为了发展趋势。图 1 给出了一个足球形 MIMO 阵列的实例。它是一种足球状的船用卫星通信 MIMO 系统，可以确保船舶在严重倾斜的情况下保持与卫星的连接。

1 MIMO 系统射频前端设计

对于任何一个大规模 MIMO 系统来说，射频前端的设计是最为核心的关键技术挑战之一。理论上讲，将已有的单路射频前端和所驱动的天线简单地并列到一起，就可以获得一个实用的 MIMO 系统，但这样简单叠加得到的大规模 MIMO 系统的成本却居高不下。这成为大规模 MIMO 系统走向商用化的第一个壁垒。

1.1 考虑成本的 MIMO 系统射频前端设计

从射频前端设计的角度看，通信系统用的 MIMO 系统与雷达用的有源

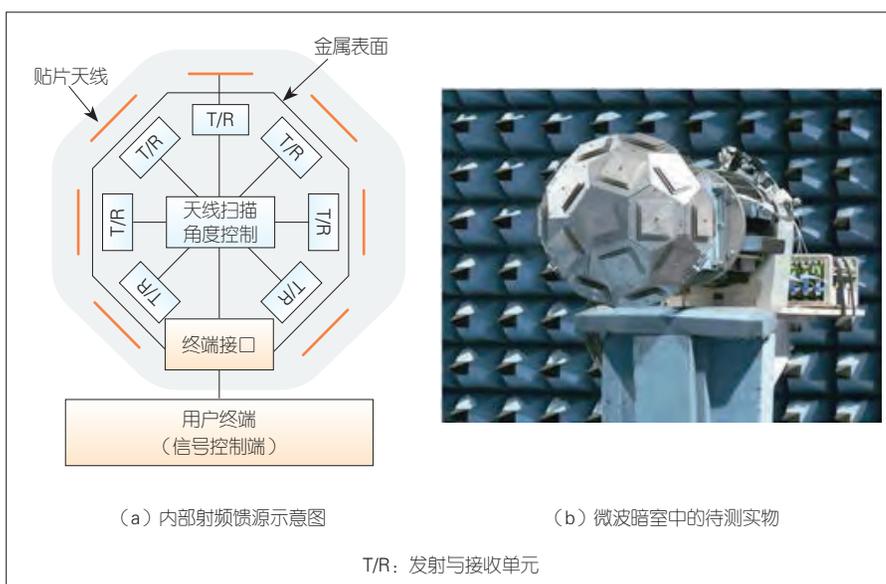
相控阵有异曲同工之处。除了信号本身以及数字基带处理不同外，两者的射频驱动电路（T/R 组件）和天线阵基本相同。与通信的大规模 MIMO 系统一样，有源相控阵雷达的目的也是实现波束扫描和波束赋形。

对于一个相控阵雷达来说，含有射频发射前端的有源天线阵部分的成本约占整个系统成本的 85%。由此类比，通信的大规模 MIMO 射频前端系统在整个基站的成本中也将占据相当大的部分。这是因为，不仅要每一个天线单元所需要的射频前端要重复出来，为了实现波束成形、波束扫描和波束控制等，还必须要增加很多额外的电路单元。为了减小甚至消除单元间的相互影响，相关的电路也需要与发射链路（Tx）等集成到一起。这些都使最后的 MIMO 系统体积加大、成本急剧增加。如何从最开始的时候就考虑到面向成本的设计（DfC），是大规模 MIMO 系统能否成功商用化的关键所在。

针对传统的单路射频系统，为了确保系统的指标，设计的时候甚至以牺牲面积为代价。比如，为了避免射

频系统中电路单元间的串扰和其他可能存在的影响，在绘制版图的时候，尽可能地将线布得比较松、电路元器件（特别是电感之间）的距离尽可能地拉大。虽然这时候的电路面积相对增大许多，但是就一路射频而言，由于这时候的设计准则是“指标优先”，这点芯片面积的增加是可以忍受的。大规模 MIMO 系统则是由很多路射频系统集成到一起的，比如，为了满足空间覆盖性和高数据率的不同要求，文献中已报道具有 1 024 路（天线单元 + 射频电路）的相控 MIMO 系统，而单载波上就已采用 256 正交振幅调制（QAM）的复杂调制形式。在电路复杂程度增加的同时，射频前端芯片的面积也在急剧增加，致使成本呈指数增长，以至阻碍了商业化。考虑到成本，在保证系统指标的前提下，应尽可能地减小射频前端的总芯片面积，因此 DfC 成为大规模 MIMO 的一个新的核心设计技术^[2]。

图 2 是一个传统射频前端的示意图，接收链路和发射链路都是独立的单向信号流走向。有时为了提高射频前端系统的性能，针对接收和发射链



▲图 1 足球状多输入多输出系统

路，也会对频率综合进行分别设计。在电路结构上，接收端的混频器和发射端的混频器仅仅是输入频率的区别。接收端的低噪声放大器和发射端的驱动功放都是放大器，并且它们所需的放大倍数也基本相同。只不过低噪声放大器对于噪声的要求比一般的功放驱动放大器要高很多，这意味着低噪声放大器完全可以用来充当发射端的驱动功放。如果能够实现驱动功放与低噪声放大器的共享、收发共享同一个混频电路和同一个频率综合，那么理论上这样一个紧凑型收发前端的芯片面积将仅是传统收发前端面积的一半，这将使系统成本大幅降低，文献中将这种紧凑型收发前端称之为双向收发前端^[3-5]（如图 3 所示）。

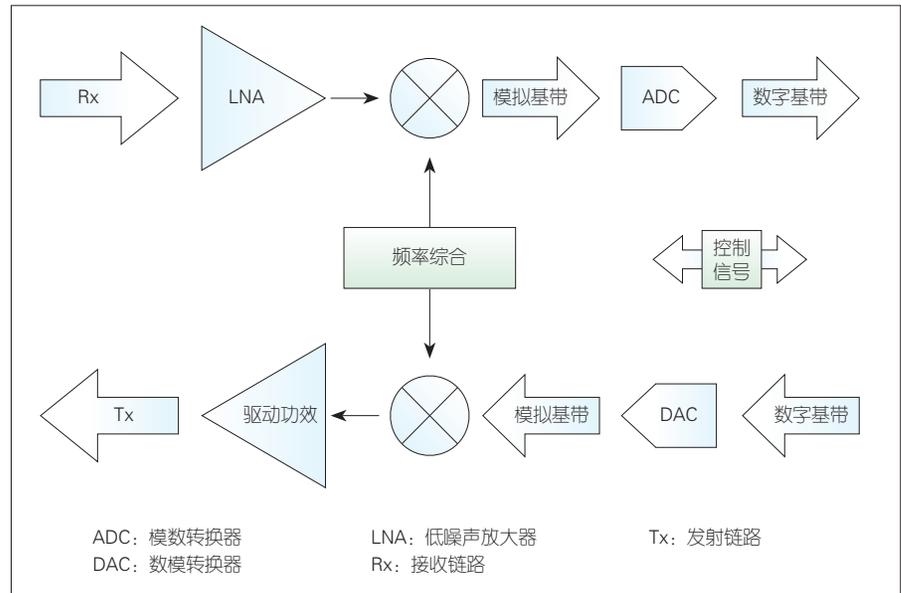
对于收发双工模式，上下行采用不同的频率同时工作。混频器本身可以同时工作在两个频率[考虑 $\cos(\omega_1+\omega_2)$ 和 $\cos(\omega_1-\omega_2)$]，这时候双向放大器也必须同时工作在两个频段；因此，双向放大器实际上就是一个双频放大器。这时上下行模拟基带同时存在，并且同时工作于同一个模拟频段，这时两个模拟基带电路无法共享。为了避免信号混淆，必须采用不同的上下行模拟基带。对于时分双工（TDD）模式则更加简单，其模拟基带电路也可以共享。

文献 [4] 利用 45 nm 互补金属氧化物半导体（CMOS）工艺实现了工作频率为 28 GHz 的高效高线性度双向收发前端芯片。文献 [5] 采用 65 nm 的 CMOS 工艺实现了满足 IEEE 802.11ay 标准的 60 GHz 双向收发前端芯片（芯片面积仅为 0.96 mm^2 ）。

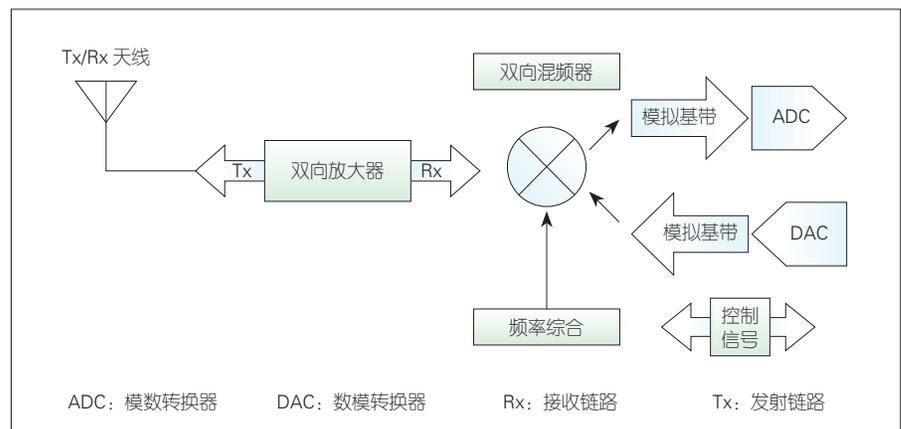
低功耗和低复杂度的优点使全射频波束赋形（aRFbf）比中频波束赋形和本振波束赋形更受到重视。全射频前端波束赋形使得利用标准硅工艺实现射频前端完全集成成为可能，这

有助于减小整个射频前端的体积和功耗，进而降低成本。更重要的是，把天线阵集成进来，在晶圆尺度实现大规模 MIMO，可使成本进一步降低^[6-7]。特别是在 100 GHz 以上的频段，其自由空间波长在 1 mm 以下。MIMO 天线阵的单元之间间隔大约为半波长。考虑到硅衬底的介电常数约 9.8，等效工作波长约为自由空间波长的一半。以 100 GHz 为例，一个标准的硅基芯片加工单元（一个光罩为 $22 \times 22 \text{ mm}^2$ ），最多可以得到 800 个单元的双极化双波束天线阵，一个标准的 12 英寸晶圆可以有 100 多个这样的标准单元。

将这些单元组装起来可以很容易获得更大规模的天线阵。比如利用 CMOS 工艺的顶层金属层来实现天线阵中的每一个天线单元、利用其他各金属层和半导体有源层来实现射频接收前端电路，这样就可以利用一个标准的 CMOS 工艺一次性实现一个完整的 MMIMO 全集成、使得成品率极大提高。避免传统上利用不同工艺平台分别实现天线阵和射频有源电路部分、然后再用封装工艺将这些系统集成所造成的很多不良因素（如采用新的衬底把用不同工艺实现的天线阵与射频前端收发 SOC 封装在一起带来的体积增大、



▲图 2 射频前端示意图



▲图 3 双向射频收发前端框图

SOC 与天线阵之间的传输损耗增大、封装带来的低成品率等), 进而带来突出的成本优势。特别地, 不同规模的 MIMO 系统都可以通过同一个工艺、同一个晶圆来实现。这样不仅能获得具有大批量一致性好的优势(同一个 MIMO 系统中各路间的一致性和作为产品的 MIMO 系统间的一致性), 而且针对不同的 MIMO 规模要求, 基于同一批次晶圆级的制造很容易实现不同的 MIMO 规模而不需要额外增加成本。基于此技术和同一批次的生产, 文献[7]在晶圆层面灵活地实现了不同规模(64 单元和 256 单元)的 60 GHz 相控阵列。

1.2 MIMO 系统超宽带相移网络设计

MIMO 系统与传统的单天线及单路射频系统的一个最大区别, 是 MIMO 系统需要实现波束赋形和波束扫描, 因此, 通过不断地调整各个天线单元之间的相位, 可以改变在远区场总的天线辐射方向。一般来说, 可以通过相移器或波束赋形网络来实现相位调整, 如 Butler 矩阵等。相移器和波束赋形网络都是与频率相关的经典电路单元。从模拟电路的角度看, 文献中有很多成熟的方案可用来获得低损耗、高精度, 甚至零静态功耗的相移网络, 但这些设计方案所得到的网络都是针对窄带工作条件的, 无法满足 5G/B5G 宽带的要求。如何使这些与频率相关的电路单元宽带化是现在面临的巨大挑战^[8-10]。全通网络(APNs)具有频率不敏感性, 自 20 世纪 90 年代起就常被用来实现宽带相移器, 文献中大多数相关的工作都是针对几吉赫兹以下频段使用分立元器件来实现的。由于这些元器件在微波低频段的品质因素 Q 值很高, 所以它们可以获得非常好的效果; 但其不足是体积过大, 同时由于短厘米波段和

毫米波端大多数分立元器件的 Q 值急剧下降, 损耗也会急剧增加。更为关键的是, 由于分立元器件的一致性较差, 在大规模 MIMO 系统中, 由分立元器件造成的相位不一致性和不确定性成为大规模 MIMO 系统的致命弱点, 这导致成品率急剧下降、成本急剧上升, 致使其无法在 5G/B5G 的大规模 MIMO 中获得有效应用。利用半导体工艺实现基于 APNs 相移网络的先天不足是其 Q 值上不去, 这导致其性能指标无法与分立元器件的实施方案相比, 但其优势是尺寸小、易于批量化生产。特别是对大规模 MIMO 系统而言, 各路间的相位一致性和确定性可以得到有效保证, 这突显了成品率优势。而在短厘米波段和毫米波频段, 半导体工艺低 Q 值的弱点不再明显(因为分立元器件的 Q 值也急剧下降)。在集成化的 APNs 相移网络成为支撑大规模 MIMO 系统商用的唯一选择的同时, 其巨大的成本优势和批量化生产加工环境, 使硅基工艺(如 SiGe、CMOS)被普遍看好^[9-10]。文献[8]实现了一个 360° 的低功耗窄带相移网络, 文献[10]利用 0.25 μm 的 SiGe 双极互补金属氧化物半导体(BiCMOS)实现了一个宽带的相移网络。

Butler 矩阵是比较经典的离散型波束赋形网络, 它通过不同耦合端口的输入激励来实现固定步长的相位调整。Butler 矩阵由一系列正交耦合器、延迟线和渐变线等构成, 这些耦合线、延迟线和渐变线等通常是由波导或微带线来实现的。波导实现的 Butler 矩阵具有高性能、低损耗的特点, 不仅可以在毫米波端来实现, 还可以拓展到太赫兹频段来实现; 但其最致命的弱点是体积大、带宽窄。相对来说, 微带实现的 Butler 矩阵比波导实现的小很多, 但损耗较大, 特别是在毫米波及以上频段, 其辐射损耗会

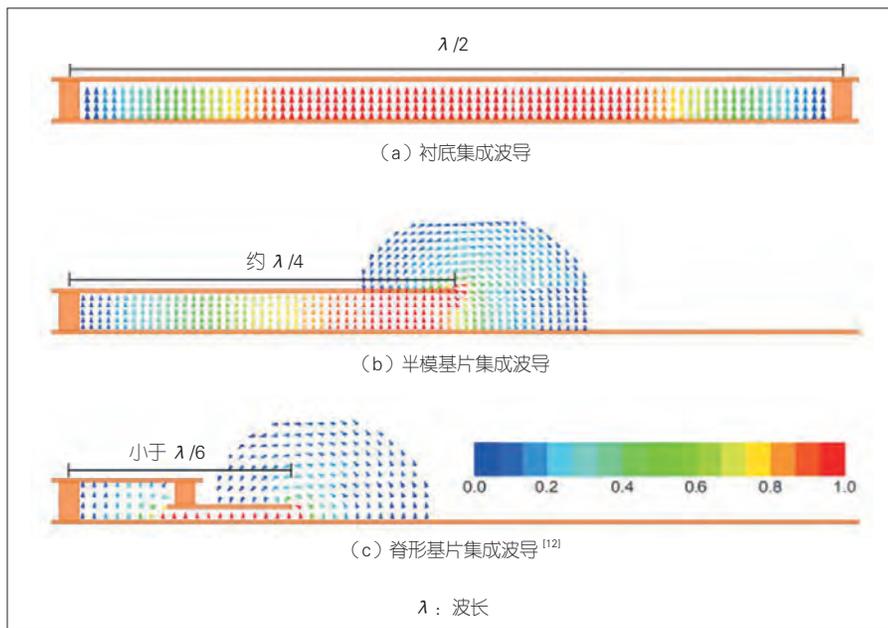
急剧增加, 导致竞争力大减。如何将 Butler 矩阵小型化且使之具有可接受的性能已成为当前一大挑战。作为一个微波无源器件小型化的技术, 衬底集成波导(SIW)在工业界已经被证明是一个低成本、小型化的技术。虽然关于在不同微波频段利用 SIW 来实现 Butler 矩阵的报道已经有很多, 但是对于大规模 MIMO 的应用来说, 其尺寸依然偏大。如何进一步使 Butler 矩阵小型化成为当前的热点课题之一。

基于传统的波导结构, 为了减小尺寸, 往往利用导波模式的对称性, 采用半模波导使实现的尺寸缩小一半。在毫米波段, 采用脊形波导可以进一步减小波导电路的尺寸。将这样的概念和做法移植到 SIW, 基于脊形半模 SIW(RHMSIW)也可实现进一步小型化的 Butler 矩阵^[11-12]。

图 4 给出了尺寸缩减的示意图。基于 RHMSIW, 文献[12]实现了一个宽带小型化的 4×4 Butler 矩阵。在保持同样的性能下, 利用 RHMSIW 实现的 4×4 Butler 矩阵比用 SIW 实现的小 70%。

将小型化的 Butler 矩阵与前面讨论的宽带相移网络集成电路结合起来, 就可以实现低成本、小型化且灵活的波束赋形, 这是一个非常有前景和实用意义的技术。

如前所述, 为满足空间覆盖性和高数据率的不同要求, 具有 1 024 个天线单元的相控 MIMO 系统已被使用, 同时单载波上也已经采用 256 QAM 的复杂调制形式。在理想情况下, 包括天线在内所有的射频通道都是完全等同的, 这也是绝大多数仿真设计所基于的前提。这样, 可以通过实现完美的波束赋形和波束扫描来达到空间覆盖性和高数据率的双重要求。众所周知, 每一个射频前端都会不可避免地使用强非线性电路单元, 特别是发射



▲图 4 归一化核横电模式场分布示意图以及相对应的尺寸比例

端的射频功放 (PA) 工作在大信号强非线性区。这带来了非线性调制, 如幅度调制 - 幅度调制 (AM-AM) 和幅度调制 - 相位调制 (AM-PM), 也使得邻近信道之间的信号泄露影响了相邻信道功率比 (ACPR) (一个描述系统线性度的重要指标)。研究表明, 在接收模式下, MIMO 若干个接收信道间的交调分量, 能够在某些方向上形成相关叠加^[13-14]。对于发射模式, 众多的 PA 很难做到完全一致, 而且 PA 内在的非线性效应也不尽相同。对一个大规模 MIMO 系统来说, 各个射频通道的非线性调制 (AM-AM 和 AM-PM) 效应几乎是准随机化的。MIMO 的阵列规模越大, 非线性调制效应随机变化就越大。最新实验表明^[14], 当各个射频通道的增益在 0.25~0.5 dB 内随机变化时, MIMO 系统的整体 ACPR 可以得到有效改善, 比如具有 256 个单元的 MIMO 的 ACPR 相对于 8 个单元的 MIMO 改善了约 3.5 dB。

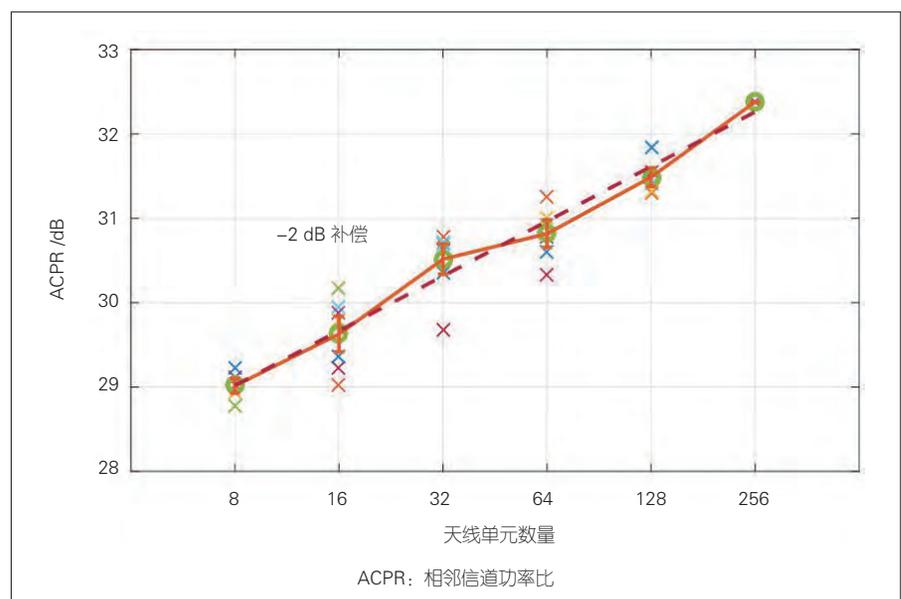
图 5 仅是实验结果, 目前尚无相关理论证明。如果这个趋势是随着阵列规模的增加而保持增加的, 相比于

8 阵列来说, 4 096 阵列的 MIMO 规模的 ACPR 的改进可能高达近 10 dB。从理论上给出相关的分析与证明对于进一步优化大规模 MIMO 性能具有重要意义。

1.3 射频本振对 MIMO 系统的影响

本振是任何一个无线通信系统射频前端都不可或缺的电路单元, 本振

的任何相位噪声都将直接影响到接收信号的矢量幅度误差 (EVM)。同样, 每一路的本振也会使任何一路发射通道的相位无法达到理论上的相位需要, 从而直接影响到 MIMO 的波束赋形和波束扫描。为了提高相位噪声指标, 传统上都采用高性能的锁相环 (PLL) 来实现低相噪。PLL 将带来很大的额外功耗 (据统计, 每一路接收通道的 PLL 功耗约为 50~200 mW, 每一路发射通道的 PLL 功耗约为 85~300 mW^[15])。也就是说, 对于一对典型的接发收通道来说, PLL 最高功耗可达 500 mW。对于一个具有 N 个单元的大规模 MIMO 来说, 仅 PLL 总功耗就高达 $N/2 W$ (对于 2 048 个单元的系统, 最高功耗可达 1 024 W)。在传统射频前端设计中, 由于只有一路, 为减小设计难度, 主要精力被放在 PLL 上而不是在本振上。但对于大规模 MIMO 系统来说, 不能简单地依靠高性能的 PLL (以功耗为代价) 来抑制本振的相噪, 因此, 对于大规模 MIMO 系统的射频前端而言, 一个设计挑战就是极低功耗本振的设计技术^[16]。根据射频 PA 的设计经验, 如果 PA 的效率很低, 那么该功放一



▲图 5 100 Mbaud 64 正交振幅调制的 ACPR 与阵列规模的关系^[14]

定会消耗更多的能量，反之，PA 的功耗就相对较小。这给了我们一个启示：如果能提高本振效率，就有可能进一步抑制本振的相噪。在高效 PA 设计中，一个很重要的提高效率的技术就是负载端的谐波控制和谐波回收。基于 0.12 μm SiGe 工艺设计，文献 [15] 利用效率提升技术实现了一款 X 波段的交叉耦合谐振电路 (LC) 本振，获得了到目前为止最好的相噪水平。

1.4 大规模 MIMO 系统的射频功放设计

虽然 5G 已经来临，但是由于受到建设周期和用户终端等因素的限制，实际上移动网络是多代混合运营的。在比较长的一段时间内，3G 和 4G 甚至 2G 都会与 5G 共存，不同的制式使用不同的射频频段。

射频 PA 是所有无线通信和移动通信中不可或缺的一个关键电路单元。从通信角度讲，人们希望 PA 可以满足所有不同频段的要求，并在理想情况下，用一个 PA 涵盖所有的工作射频频段；但从另一方面讲，PA 又是最耗能的电路单元（以 4Tx/4Rx 为例，近一半的基站能耗是由 PA 所消耗的）。因此，在设计阶段，如何实现低功耗的 PA 成为一个巨大技术挑战。众所周知，PA 的效率和带宽是矛盾的。最佳的设计是将所有的射频频段分成若干组，每一组相对宽带涵盖某些射频频段^[17-19]，因此多频率段高效 PA 设计成为当前的研究热点之一。目前有多种设计方法和技巧来实现多频率段高效 PA，例如新的匹配网络架构、可调单元、可重构调谐电路、多频谐波控制网络和其他设计技术。由于高效且可以保持一定带宽，几十年来，Doherty 功放 (DPA) 一直备受关注，同时也成为众多多频段 PA 核心基础功放的首选。标准 DPA 是一个相对窄带的功放，如何将其拓展为多频段也

成为当前的热点挑战。基于匹配网络的相位周期性和 DPA，采用阻抗变换技术和相位补偿技术，文献 [20] 实现了 6 个频率段的高效射频功放。

2 大规模 MIMO 系统射频前端设计面临的新挑战

2.1 动态空间能量分布需要新的定位理论与技术

对于传统的单收单发 (SISO) 移动通信系统来说，已知给定接发收天线位置后、空间中的能量分布是固定且可以已知的（这是利用移动通信来进行定位的前提假设），但是在大规模 MIMO 情形下，根据惠根斯原理，这些天线单元辐射出来的电磁场能量将会在空间中产生干涉现象（犹如光的干涉条纹一样）。众所周知，完全等同的多点源产生的水波存在干涉现象。由于标量水波和标量无线电波都满足同样的数学方程式——波动方程，所以两者的波形具有可类比性。在大规模 MIMO 条件下，空间中电磁波的能量分布将随点源的数量和点源的位置等的不同而具有不同的空间变化特征。当每一个独立点源的输出幅度也随时间变化时（比如，对于一个给定的大规模 MIMO 系统，根据使用环境的需求动态开启不同数量的射频通道、功放偏置电压的波动导致各路射频输出功率幅度发生波动），空间中的总电磁波能量分布就成为一个随时间和空间以及点源数量变化的函数。这些

参量是动态变化的（准随机的）、无法事先预测的。也就是说，从接收端（观测）的角度看到某一点的能量强度后，无法确定该位置所检测到的能量幅度的变化，是由单点源辐射出的功率变化导致的，还是由不同数量的多点源辐射叠加后造成的（这是与传统的 SISO 系统所不同的）。为了更好地定量描述这个现象，我们假设只有 3 个同频不同幅度的点源，它们辐射的最大功率密度分别是 $P_1=100 \text{ W/m}^2$ 、 $P_2=25 \text{ W/m}^2$ 、 $P_3=16 \text{ W/m}^2$ ，则在空间中的最大（亮斑）和最小（暗斑）能量会随着发射天线开启的数量而变化（如表 1 所示）。

对于 2G/3G/4G 来说，当给定接发收位置和传播环境时，接发收端之间的空间能量分布是固定的且可以预估的，接收端的任何能量变化都可以归结到接发收之间的距离的变化。但对于 MIMO 系统来说，空间的能量分布还会随着发射端数量的变化而变化，这使得利用传统的空间能量分布来定位变得困难。因此，必须考虑发射端数量的动态变化，这正是大规模 MIMO 带来的第一个新的挑战。

2.2 接发收端之间处于远场的假设不再完全成立

目前文献中讨论 MIMO 都基于两个基本假设：

1) 所有的天线都是点源；

2) 接发收端均处于大于 $\frac{2D^2}{\lambda}$ 的远区场 (λ 为工作波长， D 为天线的本

▼表 1 空间能量分布随发射源数量而变

开启发射端	空间能量分布 / (W/m^2)	
	最大值	最小值
$P_1+P_2+P_3$	361	1
P_1+P_2	225	25
P_1+P_3	196	36
P_2+P_3	81	1

P: 功率密度

征尺寸)。

以 N 个单元的线天线阵为例, $D=0.5N/\lambda$, 不同的线天线单元数对应的远区场条件如表 2 所示。

众所周知, 大规模 MIMO 的最有效使用场景是微小区, 也就是说, 按照点源的假设, 实际的工作场景很可能无法满足远区场的条件。

下面以 Pad 使用 WiFi 无线上网为例, 说明即使是单个天线, 点源假设也面临着挑战。假设 Pad 上的天线为点源、使用者站着上网, 通常天线距离地面(假设是理想大地)的距离约为 1.5 m。该点源的镜像点源在离大地的负 1.5 m 处, 并与真实的源构成了一对偶极子。上半空间的场分布就是这一对偶极子的辐射场, 这时 D 为 3 m, WiFi 的频率约为 2.5 GHz (波长

为 0.12 m)。此时, 远区场条件则在 150 m 之外。这意味着, WiFi 的热点只有在离使用者 150 m 之外, 才能满足通常的远区场定义, 但在实际使用中都是在十几米之内, 远没有达到通常的远区场要求。

图 6 给出了一个辐射体空间场分布的大致分类: 小于波长范围内的近区耦合场、大于 $\frac{2D^2}{\lambda}$ 的远区场、介于近区耦合场和远区场之间的近区辐射场。传统的移动通信(2G/3G/4G)和无线通信(WiFi、蓝牙等)都是以远区场作为前提来设计整个系统的, 但在实际场景中(也属于 5G/B5G 的应用场景), 除了远区场, 也有大量的场合是在近区场辐射下的。在近区场辐射条件下, 由标量电磁场假设所带来的误差不能被忽略, 这时电磁波是

一个复数矢量波(由波印廷矢量来描述)。与远区场的实数波阻抗不同, 在近区场辐射时, 波阻抗也是复数(即电场与磁场在空间和时间上都有一定的相位差, 且相位差不等于 90°)。

如果空间的电磁波能量为复数, 那么接收天线端的总能量也是一个复数^[21]:

$$P_R = P_{\text{real}} + jP_{\text{image}}, \quad (4)$$

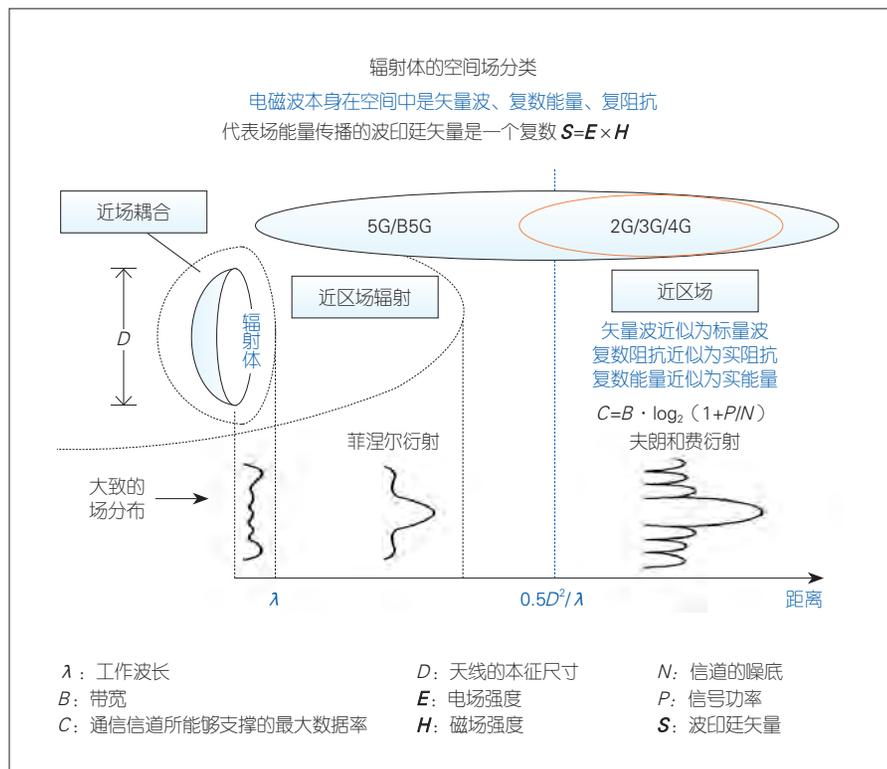
公式(4)中, $j = \sqrt{-1}$ 。由于香农定理(1)和(3)都是针对实数能量的, 对于 5G/B5G 所面临的复数矢量能量波, 如何修正香农定理给出最大信道容量成为当前一个巨大挑战。当然, 最简单的处理就是对公式(4)取模值后, 将取值代入公式(3)^[21]:

$$C_w = B \times \log_2(1 + |P_R|/N), \quad (5)$$

需要说明的是, 这里我们仅通过类比得出公式(5), 没有参照任何的数学推导或理论基础。

▼表 2 30 GHz 时不同的线天线单元数对应的远区场条件

线天线单元数 N	64	128	256	512
远区场条件 /m	20.48	81.92	327.68	655.36

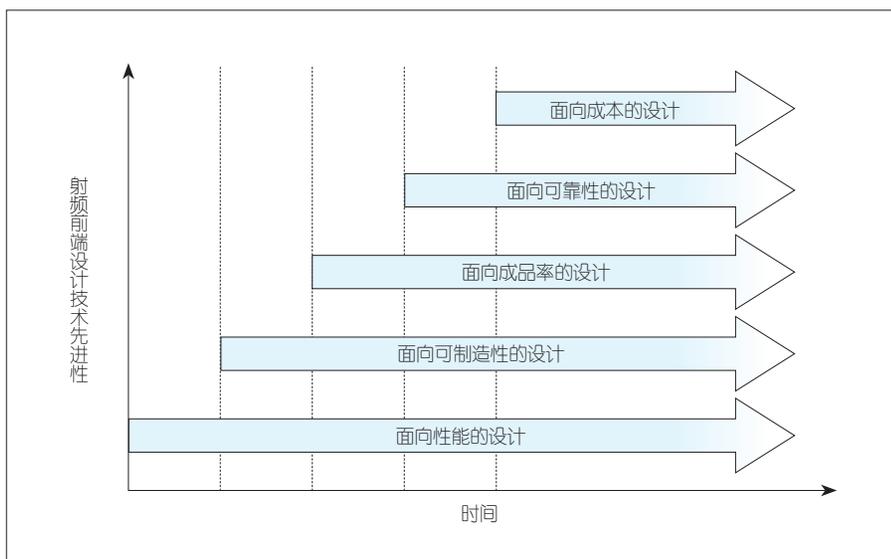


▲图 6 一个辐射体的空间场分类

3 结束语

大规模 MIMO 是保障未来网络同时满足高速率和大空间覆盖要求的技术途径, 与传统只有 SISO 的射频前端设计有着很大区别, 它不仅需要满足各项性能指标, 更为重要的是还必须考虑整个 MIMO 系统的成本; 因此, DfC 技术成为一个核心关键技术。图 7 总结了设计技术的发展趋势。

从无线电波角度看, 5G/B5G 的实际应用场景已经不再是传统的远区场。这使得广为应用的点源和远区场下实数功率、实数波阻抗、标量波等假设不再严格成立, 并且基于这些假设得到的信道最大容量必须被修正。因此, 5G/B5G 时代的一个最核心基础挑战就是如何基于矢量电磁波来研究空间无线信道传播, 而如何基于波印廷矢量来分析最大信道容量是当前面



▲图7 射频前端设计技术先进性

面临的迫切挑战。

致谢

本工作是在广东省“珠江人才计划”领军人才项目的资助下完成的，在此谨致谢意！

参考文献

- [1] SHANNON C E. A mathematical theory of communication [J]. The Bell system technical journal, 1948, (3): 379–656. DOI: 10.1002/j.1538-7305.1948.tb00917.x
- [2] MA J G. Design for cost: the key of success for 5G and beyond [J]. IEEE transactions on microwave theory and techniques, 2020, 68(1): 16. DOI: 10.1109/TMTT.2019.2961249
- [3] COHEN E, RUBERTO M, Cohen M, et al. A CMOS bidirectional 32-element phased-array transceiver at 60 GHz with LTCC antenna [J]. IEEE transactions on microwave theory and techniques, 2013, 61(3): 1359–1375
- [4] KODAK U, REBEIZ G M. Bi-directional flip-chip 28 GHz phased-array core-chip in 45 nm CMOS SOI for high-efficiency high-linearity 5G systems [J]. IEEE radio frequency integrated circuits symposium, 2017: 61–64. DOI: 10.1109/RFIC.2017.7969017
- [5] PANG J, TOKGOZ K K, MAKI S, et al. A 28.16 Gbit/s area-efficient 60 GHz CMOS bidirectional transceiver for IEEE 802.11ay [J]. IEEE transactions on microwave theory and techniques, 2020, 68(1): 252–263
- [6] MA J G. Wafer-scale all-RF beamforming phased-array transceivers for 5G and beyond [J]. IEEE transactions on microwave theory and techniques, 2020, 68(7): 2473–2474. DOI: 10.1109/TMTT.2020.3001416

- [7] KODAK U, RUPAKULA B, ZIHIR S, et al. 60 GHz 64- and 256-element dual-polarized dual-beam wafer-scale phased-array transceivers with reticle-to-reticle stitching [J]. IEEE transactions on microwave theory and techniques, 2020, 68(7): 1–23. DOI: 10.1109/TMTT.2020.2969904
- [8] MA J G. Ultra-broadband phase shifters for 5G mobile applications [J]. IEEE transactions on microwave theory and techniques, 2020, 68(2): 530. DOI: 10.1109/TMTT.2020.2965850
- [9] GARG R, NATARAJAN A S. A 28 GHz low-power phased array receiver front-end with 360 RTPS phase shift range [J]. IEEE transactions on microwave theory and techniques, 2017, 65(11): 4703–4714
- [10] ANJOS E V P, SCHREURS D M M, VANDENBOSCH G A E, et al. A 14–50 GHz phase shifter with all-pass networks for 5G mobile applications [J]. IEEE transactions on microwave theory and techniques, 2020, 68(2): 1–13. DOI: 10.1109/TMTT.2019.2948852
- [11] MA J G. Miniaturized butler matrix and tunable phase shifters for 5G and beyond [J]. IEEE transactions on microwave theory and techniques, 2020, 68(8): 3209
- [12] DER E T, JONES T R, DANESHMAND M. Miniaturized 4 × 4 Butler matrix and tunable phase shifter using ridged half-mode substrate integrated waveguide [J]. IEEE transactions on microwave theory and techniques, 2020, 68(8): 3379–3388. DOI: 10.1109/MWSYM.2019.8700857
- [13] MA J G. Overall efficiency improvements of phased arrays for 5G and beyond [J]. IEEE transactions on microwave theory and techniques, 2020, 68(3): 914. DOI: 10.1109/TMTT.2020.2972193
- [14] RUPAKULA B, ALJUHAN A H, REBEIZ G M. ACPR improvement in large phased arrays with complexmodulated waveforms [J]. IEEE transactions on microwave theory and techniques, 2020, 68(3): 1045–1053. DOI: 10.1109/TMTT.2019.2944824

- [15] WAGNER E, SHANA' A O, REBEIZ G M. A very low phase-noise transformer-coupled oscillator and PLL for 5G communications in 0.12 μm SiGe BiCMOS [J]. IEEE transactions on microwave theory and techniques, 2020, 68(4): 1–13. DOI: 10.1109/TMTT.2019.2957372
- [16] MA J G. High-performance synthesizer design for 5G and beyond [J]. IEEE transactions on microwave theory and techniques, 2020, 68(4): 1216. DOI: 10.1109/TMTT.2020.2978652
- [17] MA J G. Highly-efficient wideband RF power amplifier design for 5G and beyond [J]. IEEE transactions on microwave theory and techniques, 2020, 68(5): 1620. DOI: 10.1109/TMTT.2020.2985137
- [18] PANG J Z, LI M, ZHANG Y K, et al. Analysis and design of highly-efficient wideband RF-input sequential load modulated balanced power amplifier [J]. IEEE transactions on microwave theory and techniques, 2020, 68(5): 1741–1753. DOI: 10.1109/TMTT.2019.2963868
- [19] MA J G. Multiband RF power amplifiers for 5G and beyond [J]. IEEE transactions on microwave theory and techniques, 2020, 68(6): 2168–2171. DOI: 10.1109/TMTT.2020.2993918
- [20] PANG J Z, DAI Z J, LI Y, et al. Multiband dual-mode doherty power amplifier employing phase periodic matching network and reciprocal gate bias for 5G applications [J]. IEEE transactions on microwave theory and techniques, 2020, 68(6): 2382–2397. DOI: 10.1109/TMTT.2020.2971481
- [21] MA J G. The challenging of radio access technology for 5G [C]//IEEE International Wireless Symposium. China: IEEE, 2019: 1–4. DOI: 10.1109/IEEE-IWS.2019.8803884

作者简介



马建国，广东工业大学教授，长江学者特聘教授，百千万人才工程国家级人选、国家杰出青年科学基金获得者、IEEE Fellow、《IEEE Transactions on Microwave Theory and Techniques》主编，曾担任《Proceedings of the IEEE》编委（2013–2018）和《IEEE MWCL》副主编（2003–2005）；长期从事针对无线通信的射频电路与系统的设计和集成技术等方面的研究工作；作为项目负责人主持国家重点研发计划项目、国家科技重大专项（03专项）项目、国家自然科学基金重点项目、科技部国际合作重点项目等；发表SCI论文250余篇，获国际授权专利43项、中国授权发明专利51项。



无线物理层认证技术： 昨天、今天和明天

Wireless Physical Layer Authentication Technology: Yesterday, Today, and Tomorrow

任品毅 /REN Pinyi, 徐东阳 /XU Dongyang

(西安交通大学无线通信研究所, 中国 西安 710049)
(Institute of Wireless Communications, Xi'an Jiaotong University, Xi'an
710049, China)

DOI: 10.12142/ZTETJ.202004013

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20200722.1043.002.html>

网络出版时间: 2020-07-22

收稿日期: 2020-06-15

摘要: 提出了物理层认证技术的未来研究方向, 具体包括基于信号内生特征的无线物理层认证技术, 面向 5G 的安全、可靠、低时延无线物理层认证协议设计, 以及面向 6G 的无线物理层认证体系设计。物理层认证技术为无线网络中信息认证提供了灵活的安全保障。未来无线空口技术、网络架构和业务场景的新特性, 使得现有研究难以为未来无线认证提供全方位的安全防护, 而研发新型无线物理层认证技术在诸多方面存在挑战。

关键词: 物理层安全; 无线物理层认证; 5G; 6G

Abstract: Future research directions of physical layer authentication technology are proposed, including wireless physical layer authentication technology based on signal endogenous features, secure and reliable low-latency protocol designed for wireless physical layer authentication in 5G, and wireless physical layer authentication system designed for 6G. Physical layer authentication technology provides flexible security guarantee for information authentication in wireless networks. However, the emergence of new features of future new radio technologies, network architecture and service scenarios makes it difficult to provide all-round security protection for future wireless authentication, and there are still many challenges in developing new wireless physical layer authentication technologies.

Keywords: physical layer security; wireless physical layer authentication; 5G; 6G

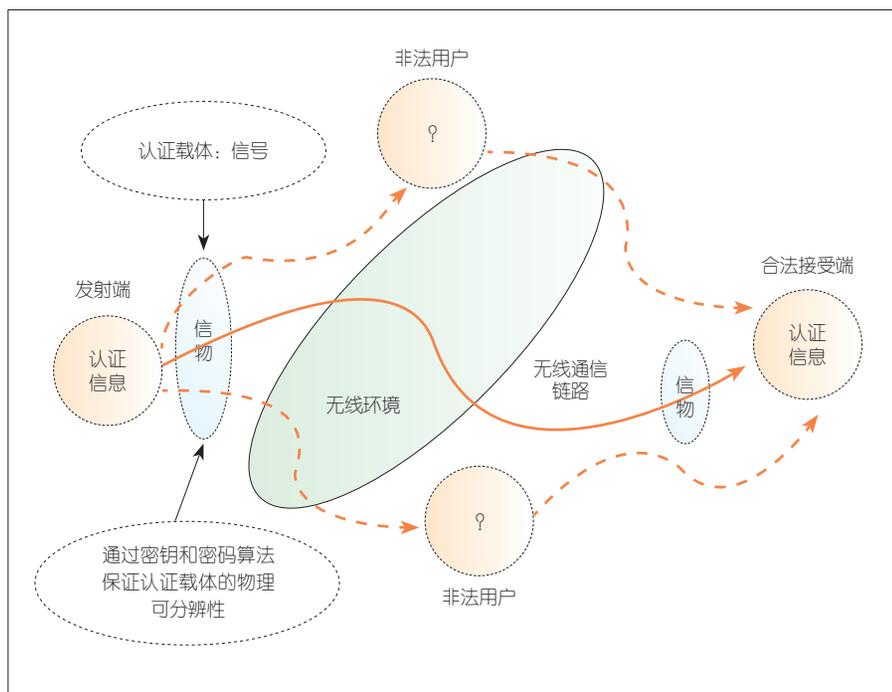
1 无线认证的起源与发展

认证是通过验证被认证对象的持有信物来证实该对象是否属实和有效, 这些信物因人而异。数字时代来临之前, 人与人之间凭关系相互识别。随着个体数量逐渐增多, 陌生人也随之增多, 人与人之间难以仅通过关系维持相互合作, 对个体进行认证成为必然, 这时承载认证信物的载体主要是语言和实物。数字时代的来临使得人与人以及人与物之间交流的方式发生变革, 载体的存储和传输更多是以数字化信息为基本方式, 认证则表现为虚拟化、数字化。

为了保障认证的安全性, 需要对载体的存储和传输方式进行加密保护。现代密码在 20 世纪 60 年代得以推出, 目的是保护私人信息免受窥探。由于密码的安全性高度依赖于个人选择, 因此其安全性十分受限。20 世纪 70 年代, 一些学者提出了公开密钥体制, 运用单向函数的数学原理, 以实现加解密密钥的分离。其中, 加密密钥是公开的, 解密密钥是保密的, 从而极大地提高了认证密钥的安全性。20 世纪 80 年代初, 美国科学家 L. LAMPOR 首次提出了利用散列函数产生一次性口令的思想, 即用户每次登录系统时使用的口令是变化的, 提高

了加密机制的安全性。20 世纪 90 年代, 美国、加拿大等国相继开展了公钥基础设施 (PKI) 的研究和建设, 为公开密钥体制提供了必要的基础设施。为了融合多种身份验证机制, 多因子身份认证 (MFA) 于 21 世纪初被提出, 为未来认证提供了基础性框架。

随着数字化时代的来临, 认证信物载体的传播方式发生变革, 从根本上改变了安全认证的模式。1897 年, 意大利科学家 G. MARCONI 首次实现了无线电波信号的远距离传输, 标志着人类进入无线通信时代。信息的无线传输导致认证无线化, 特别是认证信物载体的数字化、多样化。图 1 给



▲图1 无线认证的基本框架

出了无线认证的基本框架，发射端将认证信物嵌入认证载体（即无线信号）上，通过密钥和密码算法保证认证载体的物理可分辨性，从而使接收端识别发射端身份和信息，同时有效对抗非法用户的窃听和恶意篡改等行为。无线认证信物的载体表现为4类，包括口令特征（密码、私密密钥等）、持有特征（银行卡、密保卡等）、行为特征（语音识别、步态识别等）、生理特征（指纹识别、视网膜识别等）等。与此同时，认证无线化也引入了更多安全风险。例如，认证攻击种类繁多，包括身份假冒、数据篡改、重放攻击，以及通信抵赖。

随着无线通信与密码学不断发展以及相互融合，无线认证技术得以不断发展和完善。自从1978年1G通信诞生以来，无线认证的安全性一直是首要问题。1G几乎没有采取安全措施，移动台把其电子序列号（ESN）和网络分配的移动台识别号（MIN）以明文方式传送至网络，安全隐患极大。

20世纪90年代2G通信诞生了，但其安全机制都是基于私钥密码体制，即通过采用基于“挑战-响应”的共享秘密数据（私钥）的安全协议来实现对接入用户的认证和数据信息的保密。在此基础上，3G、4G系统对该体制进行了较大改进，但仍然是基于私钥密码体制，难以实现用户数字签名。针对4G网络认证中存在的安全问题，5G认证体系进行了修正，最典型的就是使用公私钥加密体制，增强了手机身份认证的安全性。由此可见，在5G时代，无线认证技术仍然沿用70年代的密码学原理。

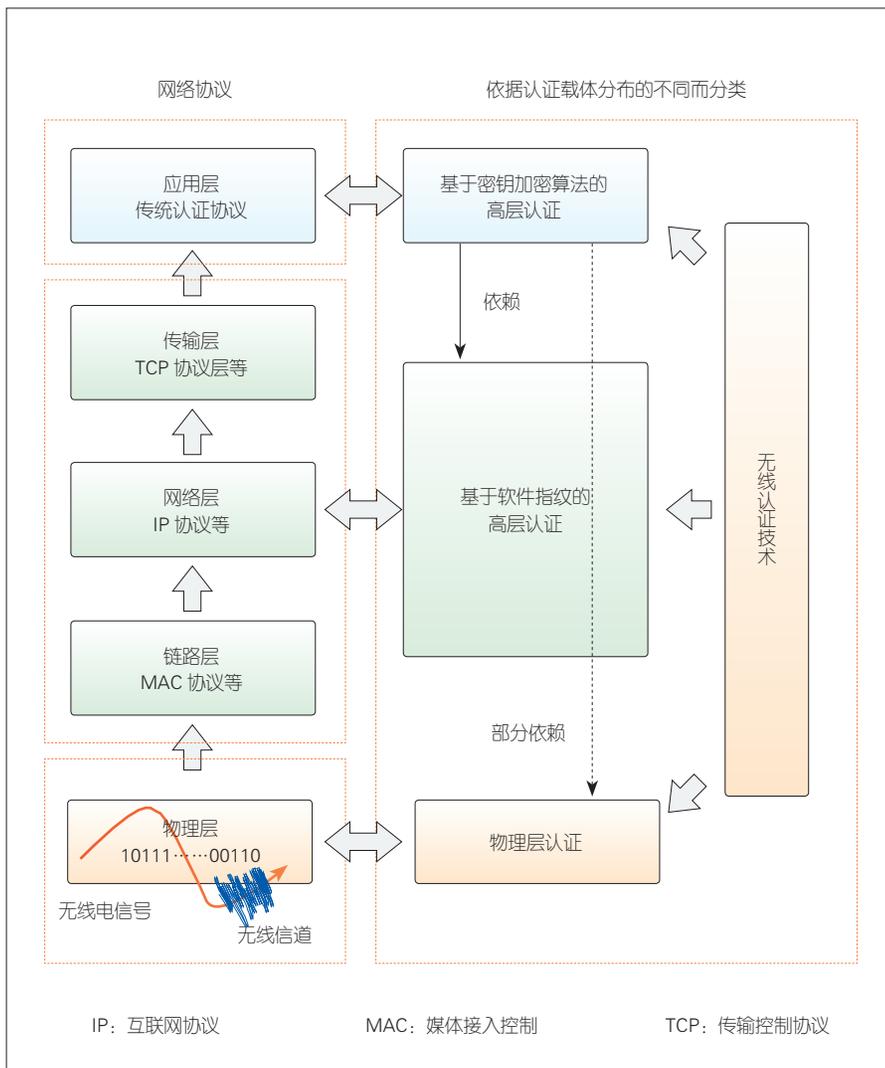
2 无线物理层认证技术及其研究现状

2.1 无线物理层认证技术的产生

认证载体是无线认证技术中最关键的部分，认证载体既承载了认证所需信物、密钥等信息，又具有多样化的表现形式，例如印章、钥匙、签名、

指纹、面部轮廓、语音声波、虹膜等。如图2所示，无线通信系统中的认证载体可大体分为3类：一类是针对密码加密算法体系的关键需求信息，主要分布于应用层面；另一类位于应用层与物理层之间，主要为协议所具备的特殊属性或部件；最后一类位于物理层，主要为与物理信号直接相关的载体，例如硬件差异（频偏、I/Q偏移）引发的特殊信号、信道状态信息、接收信号强度指示（RSSI）等。根据认证载体的类别，移动通信的认证技术可概括为3类，具体包括基于密钥加密算法的高层认证机制、基于软件指纹的高层认证机制以及物理层认证。

伴随着攻击者计算能力的提升以及先进攻击方法的产生，高层认证的安全性受到极大威胁。例如，2019年9月，谷歌公司宣告在全球首次实现“量子霸权”：其量子计算机仅用200s就完成了世界第一超算Summit用1万年的时间才能完成的计算，计算能力提高了约15亿倍。根据侧信道分析攻击的原理，攻击者可以采用时序攻击的方式，基于测量一个执行单元所需的时间，获得有用信息，这些信息可以导致密钥的泄露；攻击者通过采用功耗攻击，可以对芯片电路功耗进行分析，达到攻击及非侵入性地从设备中提取加密密钥和其他机密信息的目的。此外，随着接入增加，高层认证所需的密钥分发管理更加困难，而且网络架构的复杂异构化将导致高层认证的架构兼容性更低。基于上述背景，物理层认证技术得到广泛而深入的研究。物理层认证技术通过基于物理层的特征属性来实现对身份和消息的认证，充分利用了底层信号特征属性，因而具备与高层协议透明的优良特性。除此之外，物理层认证技术还具备较高的协议架构兼容性、较高的协议灵活性以及较低的时延等特性。



▲图2 无线认证技术研究分类

2.2 无线物理层认证的信息论基础

最早的关于无线认证的信息论研究是以基于共享私密密钥的加密机制为基础,以实现无条件安全性为目标。C. E. SHANNON 最早在文献[1]中对密钥使用和私密性的性能刻画进行了理论建模。根据 SHANNON 的理论,如果密钥长度大于信息长度,合法收发端可以通过采用“一次一密”的方法使用密钥对信息进行加密,实现信息的完美私密性。然而,关于无线认证的无条件安全性研究可分为两类:一类为基于密码学的无条件安全认证;另一类为基于窃听信道模型的无条件

安全认证。

对于第一类研究, G. J. SIMMONS 在文献[2]中最早建立了一个经典的无噪声无线认证模型:合法发射机与合法接收机共享一个密钥 K , 合法发射机发射一个经过函数 f 加密的信息 M , $M=f(K,X)$, 一个主动窃听者既可以窃听合法发射机的信息,又可以伪造或者篡改这些信息,并将新的错误信息发送至合法接收机,从而干扰合法接收机对信息的认证,最终合法接收机需要通过判断接收到的信息是否为 M 来识别其是否来自于合法发射机。针对该模型, J. CARTER 和

M. N. WEGAN 在文献[3]中证明了对于特定的信息可以实现无条件安全认证。该方案需要合法收发端根据共享私密密钥的指示,从一个大小为 B 的公共已知集合中选取双方认可的哈希函数作为加密函数 f , J. CARTER 和 WEGAN 在该文献中证明如果集合是全域的,攻击成功的概率为 $1/B$ 。U. M. MAURER 在文献[4]中证明了如果共享私密密钥不更新,非法攻击成功的概率会随着密钥使用次数的增加而提高。可以看到,以上研究并没有提及利用无线物理层信息来实现无线认证。

对于第二类研究, L. LAI 等首次在文献[5]中提出一个基于物理层信道的含噪无线认证模型,在 SIMMONS 提出模型的基础上,通过采用 A. WYNER 提出的基于窃听信道的信息传输方式,将无条件私密性的优势应用至无线认证中,实现了无条件安全认证^[6]。WYNER 窃听信道模型包含一个合法发射机、一个合法接收机和一个窃听端。其中,合法发射机的发送信息为 X , 合法接收机端合法信道输出为 Y , 窃听端窃听信道的输出 Z , 系统最大私密速率可表示为 $\max I(X; Y) - I(X; Z)$ 。关于该模型的一个重要结论是:如果合法信道质量优于窃听信道质量,那么存在服从某一分布的 X 使得最大私密速率不为零,从而保障合法收发端的安全信息传输,并且使得窃听者从接收到的信号中获得不了任何信息。借助于该优势, LAI 等人在文中得到一个重要的结论:只要保障 $\max I(X; Y) - I(X; Z)$ 大于 0,就可以在噪声信道下实现共享私密密钥的多项式次复用,使窃听者的攻击效果不会随着密钥的使用次数增加而提升。然而,实际场景中 $\max I(X; Y) - I(X; Z)$ 大于 0 这一条件并不总是成立。因此, MAURER 在文献[7]中提出了一种利用合法收发端共享的

随机信息进行密钥生成的方法，弥补了对信道质量的严格要求。基于此，很多研究关注如何利用信道特征等机制进行密钥生成。

综上所述，关于无线认证的信息论研究都需要以收发端共享私密密钥为基本前提。如果研究过程中密钥的产生过程没有利用物理层信息，则可以认为该研究属于传统的高层认证；反之，该研究则为物理层认证的一个雏形。P. YU 等在文献 [8] 中通过利用哈希函数将物理层信息与共享私密密钥耦合生成一个标签或者消息认证码，之后将信息与消息认证码经由无线信道发送，合法接收端通过利用解调后的信息生成参考消息认证码，再通过对比参考消息认证码与无线接收的消息认证码，从而完成对信息的认证和提取。YU 的研究首次为物理层认证的研究提供了一个理论模型和技术框架。针对搭线窃听信道模型下的多信息认证问题，文献 [9] 提出了一种联合多信息认证和窃听信道安全传输的物理层水印信息论模型，从理论上给出了实现多信息无条件认证安全的条件，解释了物理层水印技术的性能界。上述有关无线认证信息论方面的研究工作为无线物理层认证的理论研究奠定了坚实的基础。

2.3 无线物理层认证技术的研究现状

根据采用的认证协议架构的不同，目前对无线物理层认证技术的研究可分为两大类：第一类方案以交互式协议架构为基础；第二类方案以非交互式协议架构为基础。如表 1 所示，第一类方案的综述包括物理层水印、物理层挑战响应、跨层认证以及基于物理层密钥交换的物理层认证技术。这类方案以共享私密密钥为基础，通过采用哈希函数加密和信号处理技术实现对共享私密密钥和信号内生特征

的联合处理与利用，从而提升对合法设备信息认证的准确性。第二类方案的综述包括基于射频指纹的物理层认证技术和基于无线信道指纹的物理层认证技术。这类方案不依靠共享私密密钥，而是通过利用信号处理技术实现对信号内生特征的提取和利用，以提升对合法设备信息认证的准确性为目标。下面我们将分 6 个层面对这两类方案的研究现状进行综述，其中前 4 点都是关于第一类方案的综述，第 5、6 点是关于第二类方案的综述。

2.3.1 物理层水印

在各种物理层认证技术中，物理层水印是应用最为广泛的技术之一。在文献 [10] 中，合法发射端通过利用哈希函数加密共享私密密钥和目标信号形成标签，并将标签与无线信号叠加广播发送；合法接收端对接收到的信号进行估计并利用共享私密密钥得到参考标签，进一步通过对比参考标签与无线接收的标签，实现对目标信号的物理层认证。YU 等进一步推广该方法至多载波系统^[11]并在软件无线电（SDR）平台上验证该系统^[12]。与此不同，N. GOERGEN 等在文献 [13] 中针对认知无线电系统提出了一种信号水印方案，将无线信道状态信息作为认证信号，通过利用预共享的数字签名来判定认证信号是否属于主用户信号。在文献 [10] 中，V. KUMAR 等提出了一种基于哈希算法和收发信号设计的物理层水印方法。在文献 [14] 中，KUMAR 等提出了一种基于星座旋转

的物理层水印方法。在文献 [15] 中，Y. C. RAN 等针对物理层水印技术做了改进，通过使用随机的信道状态信息来替代共享私密密钥用于生成标签，形成了新的物理层认证方法。针对物联网设备，文献 [16] 提出了一种轻量级的物理层水印架构，通过设计轻便、高性能的共享私密密钥来保障标签的安全性以及标签与信息独立性。

2.3.2 物理层挑战响应认证

关于物理层挑战响应认证技术的研究最早源于文献 [17]。该研究可以认为是高层认证在物理层面的安全增强，其基本思想为：发送方将一个随机信号（挑战）通过无线信道广播至目的端，目的端再根据密钥对接收信号变换（响应）并反向广播至发送端。发送端已知随机信号和密钥，因此可以利用信道的唯一性和互易性抵消掉随机信号并估计出密钥，并进一步根据估计的密钥是否与预期相同来判断目的端是否合法。物理层挑战响应认证本质上是一种通过联合设计密钥和信息传输方式来实现信息认证的机制。根据密钥的物理层形式和信息传输方式的不同，物理层挑战响应认证机制得到了推广和发展。文献 [18] 将传统的物理层挑战响应认证技术延伸至中继网络场景，提出了一种新型的物理层挑战响应认证机制。该机制利用不同信道的随机性和解相关特性来实现对响应分析和对目标身份的认证。针对主动感知型信息物理系统，文献 [19] 设计了基于无线信号转发的物理挑战

▼表 1 无线物理层认证技术已有研究综述分类

模型	物理层水印	物理层挑战响应	跨层认证	物理层密钥交换	射频指纹	无线信道指纹
交互式	✓	✓	✓	✓		
非交互式					✓	✓
基于密钥	✓	✓	✓	✓		
无密钥					✓	✓

响应认证机制来应对针对信息的欺骗攻击。文献 [20] 提出了一种基于多载波信道相位随机性和互易性的物理层挑战响应认证机制，该机制通过将共享私密密钥以相位的形式嵌入到收发信号来实现设备的身份认证。针对正交频分复用 (OFDM) 系统，文献 [21] 提出了一种基于人工噪声注入的物理层挑战响应认证机制，该机制通过人工噪声掩盖合法信道的相位信息，同时创造一种人工随机性来对抗窃听器，进而实现安全的设备身份认证。

2.3.3 跨层认证

跨层认证的基本出发点是实现物理层认证与高层认证的优势互补。文献 [22] 强调了跨层信息对于认证安全的重要性，特别是跨层认证可以利用物理层信道的富散射特性、随机性、互易性和时变性来弥补高层加密体制的不足。针对 IEEE 802.11 网络，文献 [23] 提出了一种基于媒体接入控制 (MAC) 层数据包和物理层接收信号强度的抗欺骗认证方案。针对异构网络中的机器类通信 (MTC) 设备，文献 [24] 提出了一种联合射频指纹和高层认证的跨层认证方案，通过高层认证机制保障设备的合法性，以及射频指纹来鉴别认证信息的真实性。针对移动认知无线网络，文献 [25] 提出了一种联合信道射频指纹和高层认证的跨层认证方案。针对智能电网机器对机器 (M2M) 网络，文献 [26] 提出了一种双层接入认证框架，该框架通过高层认证保障无线接入过程设备的身份认证，并通过基于信道特性的物理层认证机制来保护接入信道测量，为接入数据的传输提供保障。

2.3.4 基于物理层密钥交换的物理层认证

在开放的无线接入环境中，高

层认证密钥被长期多次使用因而很容易被窃听器窃取，这会导致认证的安全性丧失。虽然系统可以通过不断的密钥更新和迭代来解决这个问题，但仍然会带来不可容忍的网络开销。物理层密钥交换技术利用随机衰落信道的内生特征 (随机性、唯一性和互易性) 作为随机共享源来生成和分发密钥，弥补了高层密钥安全性不足的问题。物理层密钥的生成不需要消耗过多计算力，其安全性不依赖于计算的复杂度，而是与无线衰落信道的物理特性有关。除此之外，物理层密钥的分发更加简单、灵活。在文献 [7] 中，MAURER 提出了一种利用共享随机信息生成认证密钥的方法，奠定了基于物理层密钥交换技术的理论基础。一个关键问题是如何获取和选择随机源，J. E. HERSHEY 在文献 [27] 中将无线信道的唯一性、互易性等内生特征转化为双方共享的随机源。在时分双工系统中，合法信道上行和下行具有相同的信道内生特征，因而合法收发端可以共享相同的信道内生特征，通过将其作为共享随机源可以产生具备无条件安全性的物理层密钥。文献 [28-30] 分别提出了利用无线信道、预编码、空间调制等技术来实现物理层密钥交换。文献 [31] 提出了一种面向带内全双工技术的物理层密钥交换方案。针对毫米波大规模多输入多输出 (MIMO) 系统，文献 [32] 提出了一种基于虚拟到达角和离开角的物理层密钥交换机制。物理层密钥交换技术可以用于替代高层密钥，进而与其他基于高层密钥的物理层认证技术相结合。例如，针对 OFDM 系统，文献 [33] 提出了一种基于物理层密钥的物理层挑战响应认证机制，其中物理层密钥从合法收发端间的信道状态信息中获取。针对时分双工 OFDM 系统，文献 [34] 提出了一种基于无线信道相位信息估计的

安全密钥生成机制来联合优化设计相位信息损失、安全密钥长度以及密钥的安全性。针对终端直通 (D2D) 中继网络，文献 [35] 则提出了一种基于社交信任和社交互易性的物理层密钥生成机制，采用博弈理论优化社交配对，从而最大化安全密钥生成速率。

2.3.5 基于射频指纹的物理层认证

上述 4 种方案合理运作的基本前提是维持合法收发端高层密钥和物理层密钥等共享私密密钥的完美私密性。与这些方案不同，基于射频指纹的物理层认证技术的核心思想是：将无线设备的硬件不完美信号特征 (射频指纹) 提取作为密钥，这些密钥因设备不同而不同，因而可以用于识别设备身份和检测非法用户；但是射频指纹数据库仍然可以被嗅探和学习，无法维持绝对的保密性。文献 [36] 验证了将该技术用于实际无线环境中鉴别无线设备身份的可行性。文献 [37] 从 OFDM IEEE 802.11a 无线信号的非瞬态前导码响应中提取双树复小波变换后的信号特征，在小波域建立了基于射频指纹的物理层认证机制。针对物联网设备，文献 [38] 提出了一种基于长短期记忆 (LSTM) 深度神经网络的射频指纹生成方法，利用无线信号的 I/Q 数据流之间的时间相关性，从大量的不完备硬件设备信号特征中训练得到可以用于识别低功率物联网设备的特征，从而保障合法设备的身份识别。针对无人机网络，文献 [39] 提出了一种基于信号能量瞬态的无人机物理层认证机制，通过能量域和时间域的信号处理技术提取无人机的信号特征，采用机器学习的方法对信号特征进行分类、识别，进而保障无人机的身份识别。文献 [40] 研究了基于频域稳态特征的射频指纹生成方法。考虑到每个设备时钟扭曲的唯一性，文

文献[41]采用时钟扭曲测量值作为设备的特征标识并将其用于身份认证。文献[42]设计了一种物理不可克隆函数，基于该函数系统可以从无线设备的微电子芯片中利用导线和晶体管的随机时延特性来生成特征标识并将其用于身份认证。针对毫米波通信，文献[43]提出了一种基于波束赋形空时模式特征的物理层认证技术方案。针对物联网设备，文献[44]将基于射频指纹的物理层认证系统建模为一个具有解析表达式的输入输出系统，从而提供了一个通用性的设计思路，该方案不依赖数据同时具备高稳健性。针对物联网设备，文献[45]则提出了一种基于多采样卷积神经网络的射频信号特征提取的方法，解决了传统射频信号特征提取过程中出现不稳定兴趣域的相关问题。

2.3.6 基于无线信道指纹的物理层认证

正如文献[8]指出的结论：认证可以看作是一个假设检验过程。通过构建二元假设检验来判断攻击的发生或者识别设备身份是另一种研究思路。基于此，基于无线信道指纹的物理层认证技术的思想是：将不同无线信道具有的多样性、唯一性和随机性特征作为一种天然的“指纹”，通过指纹的变化或者人为的指纹特征扰动构建假设检验，进而实现设备身份认证。文献[46]利用两个不同地理位置上接收机频域信道解相关的特性，通过建立一个二元假设检验过程来鉴别相干时间内两条信道所承载的信息的来源。文献[47]通过比较相邻时刻信道频率响应的变化来判断发送方是否发生了变化，进而鉴别有无攻击威胁。文献[48]利用量化的时域信道冲击响应的信号幅度和相位等信息，构建二元假设检验过程。文献[49]通过对比无线接收信号强度的差值范围，实现

移动场景中合法用户的身份认证。文献[50]通过将OFDM系统中当前时变载波偏移和偏移的预测进行对比，来实现设备身份认证。除此之外，文献[51]研究了二元假设检验过程中基于信道变化差值的自适应阈值优化方法。文献[52]通过对比不同地理位置上信号功率谱密度的差异性来实现不同位置设备的身份认证。文献[53]通过在不同无线帧之间注入人工噪声信号，使得不同时变信道下基于信号功率谱密度差异的二元假设检验更加高效，增强了身份认证的安全性。针对大规模MIMO系统，文献[54]提出了一种基于设备信道状态信息的二元假设检验，分析了不完美天线硬件特性对物理层认证机制的影响。文献[55]研究了基于极限学习机的物理层认证模型，通过联合利用无线信道的多维特征以及符合欺骗攻击模型的训练数据，提升对欺骗攻击者的安全检测性能。针对水声传感器网络，文献[56]提出了一种利用水声信道功率延迟谱，以区分不同传感器的物理层认证方案，该方案采用强化学习来选择身份认证参数，对网络和欺骗模型具备很高的透明性。针对车联网，文献[57]提出了一种用于抵御恶意边缘攻击者的物理层认证方案，该方案利用移动设备及其服务边缘共享的设备信道状态，通过强化学习、迁移学习和深度学习来达到身份认证参数选择、节省学习时间以及优化认证性能的目的。针对多用户多输入单输出OFDM系统，文献[58-60]设计了一种信号特征编码的多用户物理层认证协议，揭示了如何通过信号内生特征进行编码来实现轻量级、低时延、高安全性的多用户导频信号物理层认证。针对车联网车辆到基础设施OFDM通信系统，文献[61-62]设计了物理层Cover-Free编码理论并构建了新型的

多车辆物理层认证协议，揭示了如何在信号内生特征编码的环境下通过借助大规模天线的高空间分辨率来实现对攻击行为的精准检测、分离、识别和对攻击者的地理位置溯源，从而实现高安全、低时延的多车辆导频信号物理层认证。

3 未来无线物理层认证技术挑战

随着下一代空口技术、网络架构和业务场景的升级，研发新型无线物理层认证技术仍然是一个充满挑战的课题。

3.1 低时延物理层认证架构设计

传统的物理层认证协议大多基于交互式认证架构，随着网络接入架构的复杂化、异构化，在无线接入过程中不同交互式认证协议间切换开销急剧增加。除此之外，交互式架构下认证服务的等待时间参差不齐，在复杂传播环境下易引发过多的交互延迟。伴随着空口技术框架的革新，信号内生特征逐渐丰富，物理资源空间得到巨大扩充，为新型物理层认证协议架构的重新设计提供了更多的资源维度。然而，传统的物理层认证体系对这些特点鲜有关注。

3.2 高安全性物理层认证机制设计

传统的物理层认证协议机制依赖于共享私密密钥，在无线接入过程中，基于共享私密密钥的认证机制容易导致高交互延迟和弱计算安全性的问题；而基于物理资源空间的认证机制大都缺乏更为有效的资源信息，安全性能桎梏明显。随着下一代无线接入网络中信号与资源、协议特征的耦合性增强并表现出丰富的内生特征，用于认证的可用低维物理资源空间将得到极大的扩充。然而，传统的物理层认证体系对这些特点鲜有关注。

3.3 面向差异化安全保障能力的物理层认证协议设计

不同业务场景下具备不同安全保障能力的设备共存是下一代无线网络接入的一大特点，然而由于设备安全保障能力的差异化以及传统空口协议的固化，传统的无线接入物理层认证协议的安全性能控制相对僵化，难以保障具备不同安全保障能力的设备的安全性能。随着空口技术框架的革新，灵活的空口协议使得设备的安全保障能力得到显著提升，物理资源可以根据设备能力和安全需求进行灵活配置，因而赋予了物理资源使用和物理层认证协议设计更强的灵活性。然而，已有的物理层认证体系对这些特点鲜有关注。

4 未来物理层认证技术研究方向

经过 10 余年的发展，无线物理层认证技术得到了广泛研究和深入拓展。面向未来，无线物理层认证技术在如下几个研究方向存在巨大潜力。

4.1 基于信号内生特征的无线物理层认证技术

随着无线接入技术的革新、网络接入架构的复杂异构化以及用户接入设备数量和形态的急剧增多，空口技术、网络结构、设备能力都发生了巨大的变化。与此同时，信号内生特征逐渐丰富和多样化，不仅包括物理设备本身所具备和衍生的物理特性、与物理设备所连接的无线信道所具备和衍生的特征属性，还包括用户使用不同资源和协议时的模式特征等。然而，传统的信号特征处理方式难以深度挖掘和利用信号内生特征，表现为对信号内生特征的认知和处理能力不足，无法针对上述变化在架构、机制以及性能方面提供安全保障。实际中，无线信号从发射端经由无线信道传输至

接收端的过程中会承载和记忆诸多来自于物理设备、无线信道和使用模式的特征，包括信号的能量特性、信道随机性和独立性等。如何通过对信号内生特征进行提取和编解码使得这些特征能安全地表征和传递信息，实现信息传递的同时保障信息的可逆认证是一个很有潜力的研究方向。

4.2 面向 5G 的安全、可靠、低时延无线物理层认证协议设计

5G 空口技术框架的革新引发了对无线物理层认证协议设计新的思考。5G 系统可根据不同的场景配置多种波形技术，实现灵活自适应的空口，增强系统对各种业务的支持能力，提高系统的灵活性和可扩展性。然而，波形的设计会直接影响信号的收发和传输，产生新的信号收发模型，为无线物理层认证设计提供新的环境和思路。

在无线接入方面，5G 引入了免调度竞争接入作为备选接入方式，通过引入免调度竞争接入机制，上行传输中设备的每次传输不再根据基站的上行授权来指示，进而设备与基站间的控制信令交互大幅度降低，极大地降低空口接入时延。然而，免调度竞争接入要求用户接入、信道训练和数据识别同时进行，引发了严重的安全问题，如何在免调度竞争接入环境中设计无线信号物理层认证协议来保障接入安全具有重要意义。

在业务场景方面，作为 5G 通信 3 大应用场景之一，超可靠低时延通信 (URLLC) 对应以自动驾驶、工业控制、远程医疗以及触感网络为代表的实时关键控制类业务。URLLC 的性能指标主要包括两个部分：时延和可靠性。不论是对于时延还是可靠性，无线信道状态信息的获取都起着至关重要的作用。如果无线信道状态信息的真实性被破坏，时延和可靠性必然会降低；

因此如何设计无线物理层认证协议保障 URLLC 上行传输的信道状态信息是一项非常有意义的研究方向。

4.3 面向 6G 的无线物理层认证体系设计

未来 6G 将以 5G 的 3 大应用场景 (大带宽、海量连接、超低延迟) 为基础，实现“智慧连接”“深度连接”“全息连接”和“泛在连接”，为无线物理层认证体系的设计提供了全新的研究环境。在智慧连接方面，人工智能 (AI) 的安全性问题与 AI 技术本身相伴而生，特别是 AI 的安全识别机制，其直接影响和决定了 AI 技术的预期性能。因此，在 6G 的环境中，AI 问题将得到继承甚至强化，通过利用无线物理层认证技术可以充分挖掘无线信号的特征，从通信的角度来增强传输 AI 的安全性。在深度链接方面，6G 将关注触觉网络等方面的研究，通信设备及其连接对象将具备深度的感知、学习、实时的反馈与响应等功能。为此，6G 对低时延和高可靠的安全保障要求极高，如何利用无线物理层认证技术提取深度数据特征并且支持深度链接是一个潜在的研究方向。在全息连接方面，未来 6G 将媒体交互形式升级为全息信息交互，进而无线全息通信将成为现实。一方面，全息通信将提供更多的数据特征，为未来多因素无线物理层认证提供更多认证资源，提供更高的安全性水平。另一方面，全息通信对时延、可靠性、图像处理、智能化水平均有极高的要求，无线物理层认证技术未来有潜力满足这些要求。在泛在连接方面，由于大量不同类型的终端接入，有些终端设备能力强并具有一定的计算和存储能力，而有些终端设备甚至没有特定的硬件来安全存储身份标识及认证凭证。因此需要结合具体的业务场景，设计出灵

活的无线接入物理层认证协议以支持差异化的安全保障能力。

5 结束语

物理层认证技术为未来通信中的信息认证提供了高效可靠的保障。本文回顾了无线物理层认证技术研究的最新进展和成果。尽管现有的研究已经给出了多种安全认证策略,但是由于未来无线网络中空口技术、网络架构和业务场景的新特性,现有的研究尚难以为未来无线认证提供全方位的安全防护,因此还有大量的研究工作需要开展。此外,以下研究主题也值得关注:一是研究基于信号内生特征的无线物理层认证技术,提升物理层认证的安全性;二是研究面向5G的安全、可靠、低时延无线物理层认证协议设计;三是研究面向6G的无线物理层认证体系设计,充分挖掘6G的潜在认证资源,为未来无线接入认证提供安全保障。

参考文献

- [1] SHANNON C E. Communication theory of secrecy systems [J]. Bell system technical journal, 1949, 28(4): 656–715. DOI: 10.1002/j.1538-7305.1949.tb00928.x
- [2] SIMMONS G J. Authentication theory/coding theory [M]. Advances in cryptography. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/3-540-39568-7_32
- [3] CARTER J, WEGMAN M N. Universal classes of hash functions [J]. Journal of computer and system sciences, 1979, 18(2): 143–154. DOI: 10.1016/0022-0000(79)90044-8
- [4] MAURER U M. Authentication theory and hypothesis testing [J]. IEEE transactions on information theory, 2000, 46(4): 1350–1356. DOI: 10.1109/18.850674
- [5] LAI L F, EL GAMAL H, POOR H V. Authentication over noisy channels [J]. IEEE transactions on information theory, 2009, 55(2): 906–916. DOI: 10.1109/tit.2008.2009842
- [6] WYNER A D. The wire-tap channel [J]. Bell system technical journal, 1975, 54(8): 1355–1387. DOI: 10.1002/j.1538-7305.1975.tb02040.x
- [7] MAURER U M. Secret key agreement by public discussion from common information [J]. IEEE transactions on information theory, 1993, 39(3): 733–742. DOI: 10.1109/18.256484
- [8] YU P L, BARAS J S, SADLER B M. Physical-layer authentication [J]. IEEE transactions on information forensics and security, 2008, 3(1): 38–51. DOI: 10.1109/tifs.2007.916273
- [9] CHEN D J, ZHANG N, CHENG N, et al. Physical layer based message authentication with secure channel codes [J]. IEEE transactions on dependable and secure computing, 2019: 1. DOI: 10.1109/tdsc.2018.2846258
- [10] KUMAR V, PARK J M J, BIAN K G. PHY-layer authentication using duobinary signaling for spectrum enforcement [J]. IEEE transactions on information forensics and security, 2016, 11(5): 1027–1038. DOI: 10.1109/tifs.2016.2516904
- [11] YU P L, BARAS J S, SADLER B M. Multicarrier authentication at the physical layer [C]//2008 International Symposium on a World of Wireless, Mobile and Multimedia Networks. Newport Beach, CA, USA: IEEE, 2008: 1–6. DOI: 10.1109/wowmom.2008.4594926
- [12] VERMA G, YU P, SADLER B M. Physical layer authentication via fingerprint embedding using software-defined radios [J]. IEEE access, 2015, 3: 81–88. DOI: 10.1109/access.2015.2398734
- [13] GOERGEN N, CLANCY T C, NEWMAN T R. Physical layer authentication watermarks through synthetic channel emulation [C]//2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN). Singapore, Singapore: IEEE, 2010: 1–7. DOI: 10.1109/dyspan.2010.5457897
- [14] KUMAR V, PARK J M J, CLANCY T C, et al. PHY-layer authentication using hierarchical modulation and duo binary signaling [C]//2014 International Conference on Computing, Networking and Communications (ICNC). Honolulu, HI, USA: IEEE, 2014: 782–786. DOI: 10.1109/icnc.2014.6785436
- [15] RAN Y C, AL-SHWAILY H, TANG C Q, et al. Physical layer authentication scheme with channel based tag padding sequence [J]. IET communications, 2019, 13(12): 1776–1780. DOI: 10.1049/iet-com.2018.5749
- [16] ZHANG P C, LIU J, SHEN Y L, et al. Lightweight tag-based PHY-layer authentication for IoT devices in smart cities [J]. IEEE Internet of things journal, 2020, 7(5): 3977–3990. DOI: 10.1109/jiot.2019.2958079
- [17] SHAN D, ZENG K, XIANG W D, et al. PHY-CRAM: physical layer challenge-response authentication mechanism for wireless networks [J]. IEEE journal on selected areas in communications, 2013, 31(9): 1817–1827. DOI: 10.1109/jsac.2013.130914
- [18] DU X R, SHAN D, ZENG K, et al. Physical layer challenge-response authentication in wireless networks with relay [C]//IEEE INFOCOM 2014—IEEE Conference on Computer Communications. Toronto, ON, Canada: IEEE, 2014: 1276–1284. DOI: 10.1109/info-com.2014.6848060
- [19] SHOUKRY Y, MARTIN P, YONA Y, et al. Pycra: physical challenge-response authentication for active sensors under spoofing attacks [C]//the 22nd ACM SIGSAC Conference on Computer and Communications Security. USA: IEEE, 2015: 1004–1015. DOI: 10.1145/2810103.2813679
- [20] WU X F, YANG Z. Physical-layer authentication for multi-carrier transmission [J]. IEEE communications letters, 2015, 19(1): 74–77. DOI: 10.1109/lcomm.2014.2375191
- [21] WU X F, YANG Z, LING C, et al. Artificial-noise-aided physical layer phase challenge-response authentication for practical OFDM transmission [J]. IEEE transactions on wireless communications, 2016, 15(10): 6611–6625. DOI: 10.1109/twc.2016.2586472
- [22] MATHUR S, REZNIK A, YE C X, et al. Exploiting the physical layer for enhanced security [J]. IEEE wireless communications, 2010, 17(5): 63–70. DOI: 10.1109/mwc.2010.5601960
- [23] HAO P, WANG X B, REFAEY A. An enhanced cross-layer authentication mechanism for wireless communications based on PER and RSSI [C]//2013 13th Canadian Workshop on Information Theory. Toronto, Canada: IEEE, 2013: 44–48. DOI: 10.1109/cwit.2013.6621590
- [24] ZHAO C, HUANG L, ZHAO Y, et al. Secure machine-type communications toward LTE heterogeneous networks [J]. IEEE wireless communications, 2017, 24(1): 82–87. DOI: 10.1109/MWC.2017.1600141WC
- [25] LE T N, CHIN W L, KAO W C. Cross-layer design for primary user emulation attacks detection in mobile cognitive radio networks [J]. IEEE communications letters, 2015, 19(5): 799–802. DOI: 10.1109/lcomm.2015.2399920
- [26] CHIN W L, LIN Y H, CHEN H H. A framework of machine-to-machine authentication in smart grid: a two-layer approach [J]. IEEE communications magazine, 2016, 54(12): 102–107. DOI: 10.1109/mcom.2016.1600304cm
- [27] HERSHEY J E, HASSAN A A, YARLAGADDA R. Unconventional cryptographic keying variable management [J]. IEEE transactions on communications, 1995, 43(1): 3–6. DOI: 10.1109/26.385951
- [28] JORSWIECK E, TOMASIN S, SEZGIN A. Broadcasting into the uncertainty: authentication and confidentiality by physical-layer processing [J]. Proceedings of the IEEE, 2015, 103(10): 1702–1724. DOI: 10.1109/jproc.2015.2469602
- [29] TAHA H S, ALSUSA E. Secret key exchange using private random precoding in MIMO FDD and TDD systems [J]. IEEE transactions on vehicular technology, 2017, 66(6): 4823–4833. DOI: 10.1109/tvt.2016.2611565
- [30] TAHA H, ALSUSA E. Secret key exchange and authentication via randomized spatial modulation and phase shifting [J]. IEEE transactions on vehicular technology, 2018, 67(3): 2165–2177. DOI: 10.1109/TVT.2017.2764388
- [31] VOGT H, AWAN Z H, SEZGIN A. Secret-key generation: full-duplex versus half-duplex probing [J]. IEEE transactions on communications, 2019, 67(1): 639–652. DOI: 10.1109/tcomm.2018.2868714
- [32] JIAO L, TANG J, ZENG K. Physical layer key generation using virtual AoA and AoD of mmWave massive MIMO channel [C]//2018 IEEE Conference on Communications and Network Security (CNS). Beijing, China: IEEE, 2018: 1–9. DOI: 10.1109/cns.2018.8433175
- [33] CHOI J. A coding approach with key-channel

- randomization for physical-layer authentication [J]. IEEE transactions on information forensics and security, 2019, 14(1): 175–185. DOI:10.1109/tifs.2018.2847659
- [34] PENG Y X, WANG P, XIANG W, et al. Secret key generation based on estimated channel state information for TDD-OFDM systems over fading channels [J]. IEEE transactions on wireless communications, 2017, 16(8): 5176–5186. DOI: 10.1109/twc.2017.2706657
- [35] WAQAS M, AHMED M, LI Y, et al. Social-aware secret key generation for secure device-to-device communication via trusted and non-trusted relays [J]. IEEE transactions on wireless communications, 2018, 17(6): 3918–3930. DOI: 10.1109/twc.2018.2817607
- [36] DANEV B, CAPKUN S. Transient-based identification of wireless sensor nodes [C]//2009 International Conference on Information Processing in Sensor Networks. USA, 2009: 25–36
- [37] KLEIN R W, TEMPLE M A, MENDENHALL M J. Application of wavelet-based RF fingerprinting to enhance wireless network security [J]. Journal of communications and networks, 2009, 11(6): 544–555. DOI: 10.1109/jcn.2009.6388408
- [38] DAS R, GADRE A, ZHANG S H, et al. A deep learning approach to IoT authentication [C]//2018 IEEE International Conference on Communications (ICC). Kansas City, USA: IEEE, 2018: 1–6. DOI:10.1109/icc.2018.8422832
- [39] EZUMA M, ERDEN F, ANJINAPPA C K, et al. Micro-UAV detection and classification from RF fingerprints using machine learning techniques [C]//2019 IEEE Aerospace Conference. Big Sky, MT, USA: IEEE, 2019: 1–13. DOI: 10.1109/aero.2019.8741970
- [40] KENNEDY I O, SCANLON P, MULLANY F J, et al. Radio transmitter fingerprinting: a steady state frequency domain approach [C]//2008 IEEE 68th Vehicular Technology Conference. Calgary, Canada: IEEE, 2008: 1–5. DOI: 10.1109/vetecf.2008.291
- [41] KOHNO T, BROIDO A, CLAFFY K C. Remote physical device fingerprinting [J]. IEEE transactions on dependable and secure computing, 2005, 2(2): 93–108. DOI: 10.1109/tdsc.2005.26
- [42] SUH G E, DEVADAS S. Physical unclonable functions for device authentication and secret key generation [C]//2007 44th ACM/IEEE Design Automation Conference. San Diego, CA, USA: IEEE, 2007: 9–14. DOI: 10.1109/dac.2007.375043
- [43] BALAKRISHNAN S, GUPTA S, BHUYAN A, et al. Physical layer identification based on spatial-temporal beam features for millimeter-wave wireless networks [J]. IEEE transactions on information forensics and security, 2020, 15: 1831–1845. DOI: 10.1109/tifs.2019.2948283
- [44] ZHENG T H, SUN Z, REN K. FID: function modeling-based data-independent and channel-robust physical-layer identification [C]//IEEE INFOCOM 2019-IEEE Conference on Computer Communications. Paris, France: IEEE, 2019: 199–207. DOI: 10.1109/info-com.2019.8737597
- [45] YU J B, HU A Q, LI G Y, et al. A robust RF fingerprinting approach using multisampling convolutional neural network [J]. IEEE Internet of things journal, 2019, 6(4): 6786–6799. DOI: 10.1109/ijot.2019.2911347
- [46] XIAO L, GREENSTEIN L, MANDAYAM N, et al. Using the physical layer for wireless authentication in time-variant channels [J]. IEEE transactions on wireless communications, 2008, 7(7): 2571–2579. DOI: 10.1109/twc.2008.070194
- [47] XIAO L, GREENSTEIN L, MANDAYAM N, et al. A physical-layer technique to enhance authentication for mobile terminals [C]//2008 IEEE International Conference on Communications. Beijing, China: IEEE, 2008: 1520–1524. DOI: 10.1109/icc.2008.294
- [48] LIU F J, WANG X B, PRIMAK S L. A two dimensional quantization algorithm for CIR-based physical layer authentication [C]//2013 IEEE International Conference on Communications (ICC). Budapest, Hungary: IEEE, 2013: 4724–4728. DOI: 10.1109/icc.2013.6655319
- [49] ZENG K, GOVINDAN K, MOHAPATRA P. Non-cryptographic authentication and identification in wireless networks [J]. IEEE wireless communications, 2010, 17(5): 56–62. DOI: 10.1109/mwc.2010.5601959
- [50] HOU W K, WANG X B, CHOUINARD J Y, et al. Physical layer authentication for mobile systems with time-varying carrier frequency offsets [J]. IEEE transactions on communications, 2014, 62(5): 1658–1667. DOI: 10.1109/tcomm.2014.032914.120921
- [51] LIU J Z, REFAEY A, WANG X B, et al. Reliability enhancement for CIR-based physical layer authentication [J]. Security and communication networks, 2015, 8(4): 661–671. DOI: 10.1002/sec.1014
- [52] TUGNAIT J K. Wireless user authentication via comparison of power spectral densities [J]. IEEE journal on selected areas in communications, 2013, 31(9): 1791–1802. DOI: 10.1109/jsac.2013.130912
- [53] TUGNAIT J K. Using artificial noise to improve detection performance for wireless user authentication in time-variant channels [J]. IEEE wireless communications letters, 2014, 3(4): 377–380. DOI: 10.1109/lwc.2014.2318731
- [54] ZHANG P C, TALEB T, JIANG X H, et al. Physical layer authentication for massive MIMO systems with hardware impairments [J]. IEEE transactions on wireless communications, 2020, 19(3): 1563–1576. DOI: 10.1109/twc.2019.2955128
- [55] WANG N, JIANG T, LV S, et al. Physical-layer authentication based on extreme learning machine [J]. IEEE communications letters, 2017, 21(7): 1557–1560. DOI: 10.1109/lcomm.2017.2690437
- [56] XIAO L, SHENG G Y, WAN X Y, et al. Learning-based PHY-layer authentication for underwater sensor networks [J]. IEEE communications letters, 2019, 23(1): 60–63. DOI: 10.1109/lcomm.2018.2877317
- [57] LU X Z, XIAO L, XU T W, et al. Reinforcement learning based PHY authentication for VANETs [J]. IEEE transactions on vehicular technology, 2020, 69(3): 3068–3079. DOI: 10.1109/tvt.2020.2967026
- [58] XU D Y, REN P Y, RITCEY J A. Independence-checking coding for OFDM channel training authentication: protocol design, security, stability, and tradeoff analysis [J]. IEEE transactions on information forensics and security, 2019, 14(2): 387–402. DOI: 10.1109/tifs.2018.2850334
- [59] XU D Y, REN P Y, RITCEY J A. Code-frequency block group coding for anti-spoofing pilot authentication in multi-antenna OFDM systems [J]. IEEE transactions on information forensics and security, 2018, 13(7): 1778–1793. DOI: 10.1109/TIFS.2018.2800696
- [60] XU D Y, REN P Y, RITCEY J A. Hierarchical 2-D feature coding for secure pilot authentication in multi-user multi-antenna OFDM systems: a reliability bound contraction perspective [J]. IEEE transactions on information forensics and security, 2019, 14(3): 592–607. DOI: 10.1109/TIFS.2018.2859585
- [61] XU D Y, REN P Y, RITCEY J A. PHY-layer cover-free coding for wireless pilot authentication in IoV communications: protocol design and ultra-security proof [J]. IEEE Internet of things journal, 2019, 6(1): 171–187. DOI: 10.1109/ijot.2018.2878333
- [62] XU D Y, REN P Y, RITCEY J A. Reliability and accessibility of low-latency V2I channel training protocol using cover-free coding: win-win or tradeoff? [J]. IEEE transactions on vehicular technology, 2019, 68(3): 2294–2305. DOI: 10.1109/tvt.2019.2891295

作者简介



任品毅，西安交通大学教授，无线通信研究所所长；主要研究领域为无线物理层安全传输、5G与网络、认知无线网络、卫星通信与组网、信号检测、分布式网络等；先后主持和参与30余项国家级课题，在“十一五”期间被聘为国家高技术研究发展计划（“863”计划）“频谱共享无线通信系统”重点项目总体专家组副组长；发表论文150余篇，出版译著10余本，以第一发明人获国家发明专利30余项，登记国家计算机软件著作权7项。



徐东阳，西安交通大学讲师；主要研究领域为无线物理层认证技术、5G、编码理论等；2017年获《China Communications》首届最佳论文奖；发表论文20余篇。



确定性网络技术及应用场景研究

Deterministic Networking Technology and Scenarios

魏月华 /WEI Yuehua

喻敬海 /YU Jinghai

罗鉴 /LUO Jian

(中兴通讯股份有限公司, 中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTETJ.202004014

网络出版地址: <https://kns.cnki.net/KCMS/detail/34.1228.TN.20200224.1105.004.html>

网络出版日期: 2020-02-24

收稿日期: 2020-02-10

摘要: 结合多种网络场景, 探讨了确定性网络技术的实现原理。与现有技术进行对比研究, 提出了原型系统的设计方案。认为确定性网络技术是以以太网和IP网络技术由“尽力而为”向“确定性”发展的新阶段, 可以提供超低的丢包率和可控的端到端时延, 以助力运营技术(OT)和IT的融合, 有效提升网络可用性, 并显著降低网络成本。

关键词: 确定性网络; 丢包率; 时延

Abstract: The implementation principles of deterministic network technology in combination with various network scenarios are discussed. Compared with the existing technologies, the design scheme of a prototype system is proposed. It's considered that deterministic network technology is a new stage in the development of Ethernet and IP network technology, which can provide ultra-low packet loss rate and bounded end-to-end latency, help the integration of operation technology (OT) and IT, effectively improve network availability, and significantly reduce network costs.

Keywords: deterministic network; data loss rate; latency

1 确定性网络的技术特性

1.1 确定性网络的技术概览

确定性网络^[1]是为确定性业务流提供服务的网络。它不限定特定拓扑且不限制连通性, 应用时可以通过网管系统、应用控制器等来进行网络资源预留。对确定性网络感兴趣的许多应用终端, 通常需要能够支持时钟同步至亚微秒级的精度。确定性网络中的一些队列控制技术也需要中继节点和传输节点之间的时间同步。

相对于非确定性网络, 确定性网络的最核心特征是更严格、更明确的服务质量(QoS), 包括: 从源到目标的最小和最大端到端时延和时延抖动, 在节点和链路的各种假设操作状态下

的丢包率, 数据包发生乱序(比率)的上限。

目前, 已经有方法实现可控的时延和丢包来满足很多应用, 例如基于优先级和冗余配置的技术, 然而, 这些技术通常只有在关键流在网络容量中占比很小、网络中的所有系统都运行正常、没有终端系统中断网络操作行为等情况下, 才能工作得很好。确定性网络关心的是端到端延迟在最糟糕情况下的值。平均值或典型值对确定性网络没有意义, 因为它们不代表一个实时系统执行任务的能力。一般来说, 一个普通的基于优先级的队列方案比确定性网络数据流有更好的平均延迟, 但在最坏情况下的延迟却可能是不受控的。确定性网络采用拥塞保护、显式路由、服务保护来提供QoS。

1.2 确定性网络堆栈模型

图1是一个概念性的确定性网络数据面层次模型。图中的“源”和“目的”为应用层。“报文定序”“副本消除”“流复制”“流合并”“报文编码”“报文解码”均是确定性网络服务层的一部分。其中, 报文定序为报文复制和副本消除提供顺序号。如果确定性网络流由更高层的传输协议来执行报文定序和副本消除, 那么就不需要这一层。副本消除基于报文定序功能提供的序列号, 丢弃确定性网络流复制所产生的任何报文副本。副本消除功能还可以对报文重新排序, 以便从因报文丢失而中断的流中恢复报文顺序。流复制将属于确定性网络复合流的报文复制到多条确定性网络成员流里。该功能与报文定序是分离的。流复制

可以是对数据包的显式复制和重标记，也可以通过例如类似于普通多播复制的技术来实现。流合并将属于特定确定性网络复合流的成员流合并在一起。确定性网络流合并与数据包定序、副本消除、确定性网络流复制一起执行报文复制和消除。报文编码可以替代报文定序和流复制的功能。报文编码将多个确定性网络报文中的信息进行组合（这些信息可能来自不同的确定性网络复合流），并将这些信息在不同的确定性网络成员流上用报文进行发送。报文解码可以替代报文合并和副本消除的功能，并从不同的确定性网络成员流中取得报文，然后从这些报文中计算出原始的确定性网络报文。

确定性网络在传输层提供“拥塞保护”。实际的队列和整形机制通常由下层的子网层提供。“显式路由”通过确定性网络传输层提供机制，确保为确定性网络流提供固定的路径。

操作、管理和维护（OAM）可以利用带内和带外信令验证服务在 QoS 约束下的有效性，并可以在数据包中添加特定的标记，以追踪网络运行、传递或发生错误。OAM 未在图 1 中画出，因为它可以存在于任意功能层中。

1.3 确定性网络数据平面

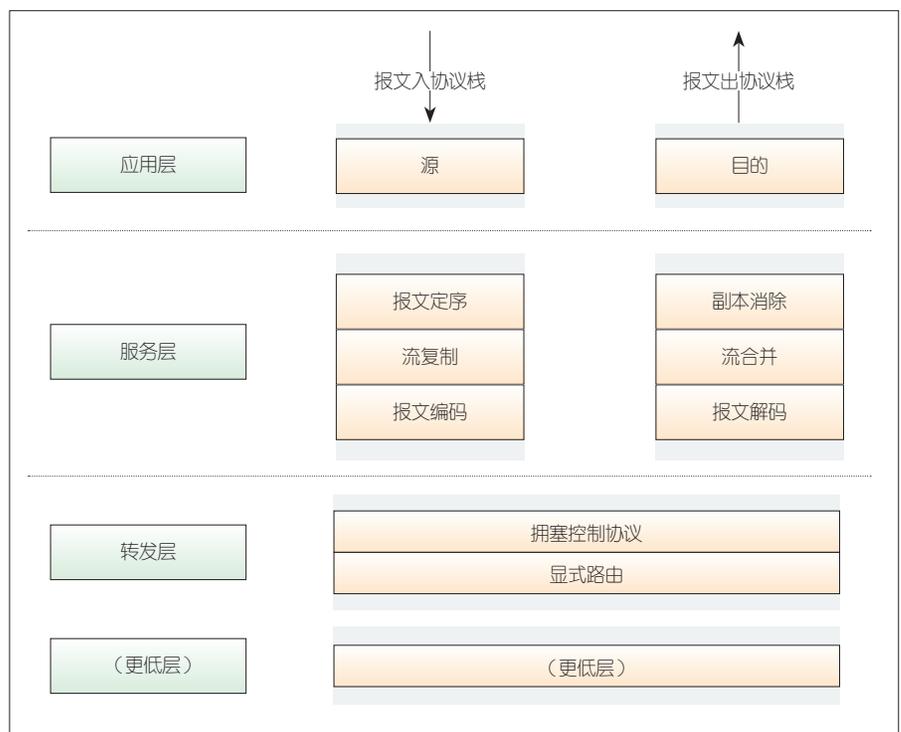
确定性网络由确定性网络使能的终端和节点组成。所有启用确定性网络的节点都连接到子网，其中点对点链接也被认为是简单的子网。这些子网提供确定性网络兼容服务，以支持确定性业务流。子网的例子包括 IEEE 802.1 时间敏感网络（TSN）和光传送网（OTN）。多层确定性网络系统也是可能的，例如将其中一个确定性网络作为子网为更高层的确定性网络系统提供服务。一个简单的确定性网络概念如图 2 所示。

分别为确定性网络服务层、传输

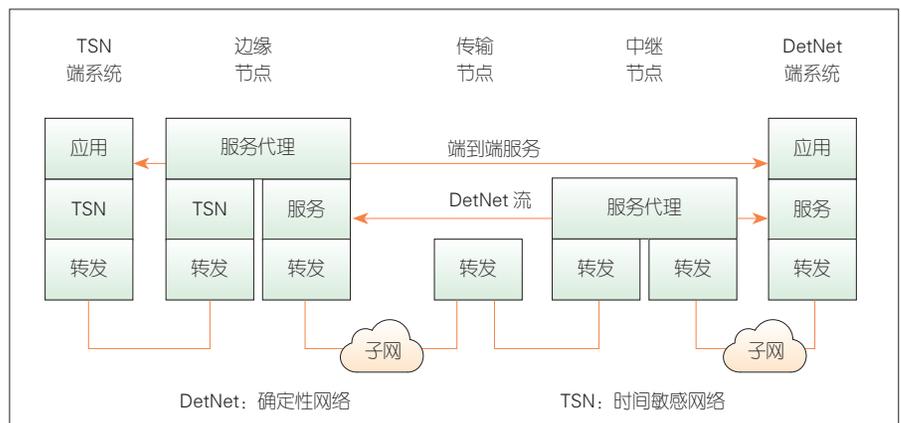
层选择一种技术方法，就可以为确定性网络流提供多种数据平面解决方案。不同的数据平面选项之间最根本的区别，是确定性网络端点系统使用的基本寻址方法和报文头各不相同。例如，可以基于多协议标签交换（MPLS）标签或 IP 报头来递送基本服务。传输层的技术选择会影响确定性网络服务层的基本转发逻辑。在这两种情况下，确定性网络节点都用 IP 地址来标示。所选的确定性网络传输层技术也需要

映射到用于互连确定性网络节点的子网技术，例如，确定性网络需要映射到 TSN 帧。

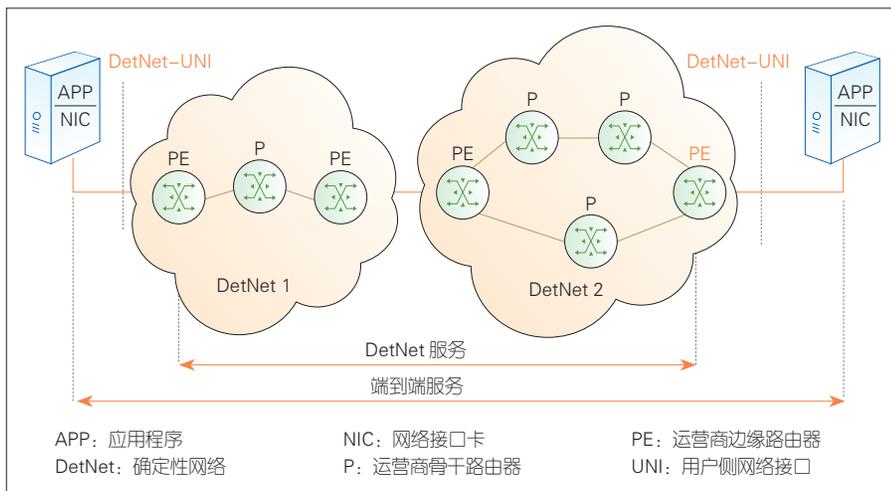
确定性网络用户侧网络接口（UNI）是基于分组的参考点，终端和 PE 通过分组网络提供连接，如图 3 所示。确定性网络 UNI 具有多种功能，例如，它可以将特定的联网技术专用封装添加到确定性网络流中，提供与预留相关的连接可用性状态，为终端系统提供同步服务。



▲图 1 数据面层次模型



▲图 2 确定性网络的使能网络组件



▲图3 确定性网络参考架构

1.4 确定性 QoS 的实现机制

1) 拥塞保护机制。该机制通过对确定性网络流所经过的路径进行预留资源来实现。预留的资源可能是缓存空间、链路带宽等。拥塞保护能极大地减少甚至完全消除网络中因输出报文拥塞造成的报文丢失，但它只作用于限定了最大报文大小和传输速率的确定性业务流。拥塞保护牵涉到确定性网络 QoS 的两个需求：延迟和丢包。鉴于确定性网络节点的缓冲区是有限的，拥塞保护必然导致最大的端到端延迟。缓冲区拥塞也是对丢包影响较大的因素。

2) 服务保护机制。除了拥塞，随机媒体错误和设备失效对丢包影响也很大。确定性网络使用报文复制和消除机制来解决这类丢包问题，从而实现服务保护。这种机制有时需要将确定性业务流重新编码并分发到多条路径上，使某条或某些路径的失效不会导致任何报文丢失。

3) 显式路由。显式路由通常通过特定的协议或者集中控制单元，根据确定性业务特性及网络约束条件计算出最佳确定性路径。这些确定性路径通常不会因路由或桥接协议的收敛而发生改变。

这 3 种机制可以独立或组合应用（有 8 种可能的组合），例如，在 IEEE 802.1CB 中采用显式路由和服务保护来实现无缝冗余机制。显式路由通过限制网络的物理拓扑为一个环来实现。顺序化、复制和副本消除是通过在以太网帧的头部或尾部加上报文标签来实现的。IEEE 802.1Qat、IEEE 802.1Qca 可以提供拥塞保护。只要网络不失效，就可使用流预留协议（SPR）或者路径控制与预留协议（PCR）在每个交换节点上做整形。如果将 3 种机制结合起来使用则低时延业务的可靠性可以获得最大程度的保障。

确定性网络通过在确定性网络流路径的每一跳预留带宽和缓冲区资源，实现拥塞保护和有限的传送延迟，然而在穿越多供应商网络时，预留本身是不够的。这是因为一个系统中的时延变化，会导致下一跳系统需要额外的缓冲空间，从而增加最坏情况下每跳的延迟。

标准排队和传输选择算法允许中央控制器计算每个传输节点对端到端延迟的贡献，和每个传输节点中每个增量确定性网络流所需的缓冲区空间量。IEEE 802 已经规定（并且正在制订）一组排队、整形和调度算法，使每个

传输节点（网桥或路由器）和 / 或中央控制器能够计算这些值。这些算法包括基于信用的整形器^[2]、基于时间同步的时间门控队列调度^[2]、基于时间同步的双重（或三重）循环队列转发调度^[2]，以及基于优先级抢占的传输机制^[2-3]。除了分组抢占技术外，它们都是可以应用到其他非以太网的媒介上。

2 应用场景举例分析

2.1 交通领域

2.1.1 汽车车内控制及娱乐通信网络

汽车领域的应用包括汽车内部的音视频和总线，也包括自动驾驶网络。前者是应用以太网局域网技术，后者是应用 IP 互联网技术。

汽车控制网络的主要特点有：高度工程化、固定拓扑、物理规模很小、通常在 30 m 五跳之内，但是会有很多端口（可能达到 100 个设备）。网络需要连接控制器、传感器、驾驶辅助视频、雷达以及娱乐用的影音，还需要通过网关连接控制器的局域网技术（CAN）、FlexRay、面向媒体的系统传输（MOST）等。

传统车载娱乐网络的主要技术是 MOST。MOST 技术的专有性质妨碍了自身应用。如果没有更多的开放性和更高的承受能力，MOST 技术可能会让位给以太网。

汽车控制网络应用需要确定性的极小时延，例如备用或驾驶辅助相机。如果使用 100 MB 的物理层（PHY），那么每跳的时延需要小于 20 us。

车上控制回路系统中的传感器和控制消息都是预先安排调度的。调度循环周期一般在 30 us~10 ms 之间，大部分情况是 125 us。控制消息一般在 128~256 字节之间。传感器需要的带

宽变量范围比较大。已有的应用所需带宽比较低,但是新型机器视觉应用在本地区域中需要更高的带宽。

IEEE 802.1Qav 排队和转发协议通过网络调度高优先级流量,确保低优先级数据不会干扰时间敏感内容。带宽保留在流启动之前保留整个网络的端到端带宽可用性,保证带宽直到明确释放。带宽预留可以预先配置为最小启动。为预期流量模式配置静态预留,从而保留所需网络资源的系统功能是汽车用例中的默认功能。

确定性网络协议作为更开放的标准,可用于下一代信息娱乐和驾驶员辅助系统的解决方案。

2.1.2 轨道交通控制及娱乐通信网络^[4]

随着轨道交通的发展,自动化、安全、舒适、娱乐需求的提升对轨道交通的通信系统提出新的要求。轨道交通中通信系统提升来自两个驱动力:一个是旅客信息系统及旅客外部网络接入的需求,二是机车自动化的需求。机车自动化又分为机车控制和机车操作维护。表1为轨道交通各种信号的网络需求。

这些参数中,机车控制信号优先级最高,机车操作维护次之,旅客信息服务优先级最低。实际上,高优先级的控制信号会被低优先级的娱乐、上网信息淹没。当前的解决方案是各种信号分开采用专用系统,但这将会

带来系统割裂、价格高昂、维护困难等问题。

能够在一个统一、标准的通信系统内共存,又能够满足各种信号的需求,是最优的解决方案。确定性网络技术正是这样的解决方案。

2.2 工业领域

2.2.1 电力传输与保护系统

电力设施部署依赖于下层网络的高可用性和行为确定性。在电力传输中,传输保护是一个非常重要的需求。电力保护包括操作者、电力设备的保护,以及电网的稳定性和频率的保持。如果出现错误,将会对操作者、电力设备和电网本身造成损害并导致断电。通信链路结合保护中继,可在最短时间内发出命令信号,以切断高压线路上的错误部分。

当前,电力传输与保护还依赖于在复杂环境下采用的技术,这包括时分复用(TDM)网络技术和一些应用特定网络技术。这种网络环境无法将OT和IT集成到同一个网络中,反而会产生信息孤岛。未来电力设施会向集成的基于开放和标准化的IP基础设施发展。

2.2.2 楼宇自动化系统

在典型的楼宇自动化系统(BAS)架构中,管理网络采用基于IP的通信

协议。在中大型楼宇中,管理系统部署在楼宇中;对于小型办公室或住宅,管理系统布放在远程,以节约成本。现场网络主要采用非IP的通信协议。本地控制器(LC)连接几十或上百个使用“现场协议”的设备,这些设备包括环境监控器、火源探测器、反馈控制等。LC一般是一个可编程逻辑控制器,负责测量设备状态,以提供信息给楼宇管理服务器或人机接口,还负责发送控制指令给设备(单方面的,或作为控制环路的反馈)。BAS中的“现场协议”五花八门,有多种介质接入控制(MAC)/PHY模块和接口,这导致BAS比较昂贵,存在很多厂商锁定的管理应用。

管理网络通常是尽力而为的,但是现场网络采用的是非IP的设计,导致无法互通。现场网络有特定的时间同步、定时等要求,未来BAS将可以提供更复杂和精确的管理控制,而且楼宇网络能够连接到企业网、家庭网和互联网。融合的网络可以代替现场网络中的现有技术。新的BAS网络架构和技术应该要保证低通信时延、低抖动、“6个9”的可靠性、容灾备份以及设备和网管之间的鉴权和认证。

2.2.3 专业音频和视频传输网络

专业音频和视频行业包括音乐和电影内容创作、广播、电影制作、现场声音、大型场馆的公共广播。这些

▼ 表1 轨道交通的网络通信需求

服务	上下行	每节点带宽需求 / (Mbit/s)	优先级	端到端时延	可靠性	安全	数据完整性	每车厢节点数	总带宽需求 / (Mbit/s)
控制信号	下行, 上行	<1	最高 (=1)	<100 ms	>99.999%	最高	强制	2	2
实时视频监控	上行	>4	高 (=2)	<500 ms	>99.99%	高	推荐	20	80
乘客信息系统	下行	<1	低 (=4)	<1 s	<99.99%	中	-	6	3
紧急语音通知	下行, 上行	<1	高 (=2)	<200 ms	>99.99%	高	-	8	6
乘客上网	下行	≥ 0.2	低 (=4)	<10 s	<99.9%	中	-	500	100
机车检测信号	上行	>0.1	中 (=3)	<1 s	<99.99%	高	推荐	50	5

行业已经将音频和视频信号从模拟转换为数字，但是数字互联系统主要是点对点的，每条链路上只有一个（或少量）信号与专用硬件相互连接。专业音频和视频的确定性网络的典型用例包括不中断的流播放、同步的流播放和声音增强。

如今专业音频和视频行业正急需基于数据包的基础设施，其应用程序可以基于 IEEE 802.1 TSN 标准，创建和传输确定性的流，但是不能通过 IP 路由，因此不能有效地分布到更广阔的区域（例如跨越广阔地理区域的广播事件）。如果这样的局域网流可以跨越 IP 路由网进行连接，则可以为专业音频和视频应用提供更为灵活的网络解决方案。

2.3 电信领域

3GPP 定义的典型蜂窝网络架构包括前传网、中传网和回传网 3 个网段。前传网连接基站处理单元（BBU）到远端射频头（RRH），中传网将基站互连起来，回传网将无线基站连接到网络控制器或网关。为了适应移动互联网和物联网的业务场景，低时延、高可靠是 5G 的 4 大关键技术特征之一。

留给前传网络的总体传输时间受限于基带处理无线帧后的可用时间。对基于分组的传输，分配的传输时间需要供天线与基带处理单元之间所有节点和缓存加上线路延迟使用。在当今的前传网网络技术中，队列、调度和发送组件成为主要因素，链路时延反而不是主要因素了，因为前传网链路相对来说很短。

一般来说，前传距离是一个给定参数，因为 RRH 和 BBU 通常放置在预定地点。然而，一个 RRH 可能会增加更多的天线，来提升多输入多输出（MIMO）容量或支持大规模 MIMO。这意味着增加共享相同的前传链路的

前传流量。确定性网络可以控制前传链路的带宽分配和流量调度，并提供足够的缓冲区来减少丢包率。

对于中传网，时延约束主要受站点之间的无线功能驱动，例如协同多点处理（CoMP）。CoMP 的设计原则是将当前一个蜂窝向多个用户终端（UE）发射的模式，扩展为通过基站站间的协作把多个蜂窝向多个 UE 发射的模式。CoMP 的“中传时延”和“信道状态信息（CSI）报告和精度”是两个时延敏感的性能参数。CoMP 的基本特征是在演进型基站（eNB）之间互访信令，所以中传网的时延是 CoMP 性能最主要的限制。站点间的 CoMP 是 5G 的关键需求。

确定性网络技术通过控制和减少队列、调度和传输操作所需要的时间，从而为链路传输留出更多时间以支持更长的传输距离。通过提供与尽力而为流量的隔离，确定性网络技术还可以满足 5G 传输网的不同网络切片差异化的性能需求。

中传和回传网络已经向着支持精

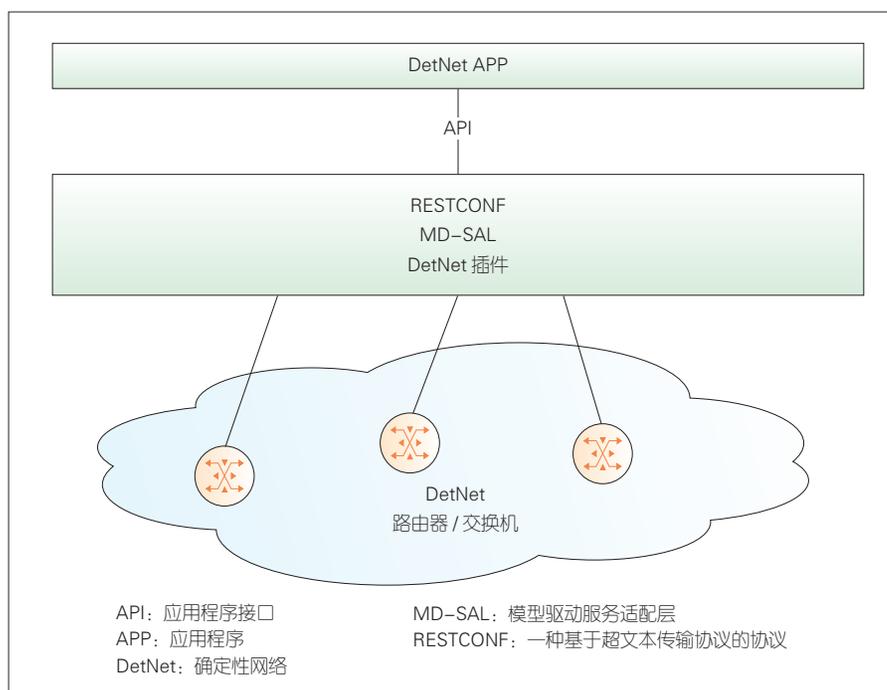
确时间同步的传输网络发展。传输网络本身为了满足带宽和成本的需求，事实上已经过渡到全 IP 的基于分组的网络，因此如此实现高精度的时钟分发已经成为一个挑战。

3 集中式确定性网络原型系统

如图 4 所示，典型的集中式确定性网络原型系统包括 3 个部分：确定性网络（DetNet）应用程序（APP）、DetNet 控制器、DetNet 路由器/交换机。

DetNet APP 向用户提供可视化界面，以方便用户对网络进行配置；DetNet 控制器基于 OpenDayLight，提供 DetNet 管理插件，主要负责拓扑收集、业务部署、路径计算、资源预留、操作维护管理（OAM）等工作，并通过网络配置（NETCONF）/边界网关协议 - 链路状态（BGP-LS）等南向接口与设备侧进行交互；DetNet 路由器/交换机执行转发面的工作，根据控制器的指令进行转发。

控制器的主要功能模块如图 5 所示。第 1 层静态处理包括 4 大功能模



▲图 4 确定性网络原型系统架构

块：“拓扑管理”模块负责拓扑收集，并获取相关链路属性（链路度量值、带宽、时延等）以及各网络设备类型和能力，提供北向接口，支持手动配置拓扑和链路属性，存储拓扑信息；“流特征管理”为动态处理单元提供确定性业务数据支撑；“域划分”模块划分 TSN 域或 Detnet 域等，并进行各域的特性数据管理；“时间同步”模块提供北向接口用于时间同步参数配置，调用南向接口插件下发配置至网络设备。

第2层次动态处理包括两大功能：“路径计算”模块根据拓扑以及业务约束条件（带宽和时延），计算满足需求的路径，供端到端业务整合模块使用；“端到端业务整合”模块调用路径计算单元（PCE）模块获得业务路径，然后进行端到端业务的部署——根据路径形成转发表来处理 TSN 域内业务，利用传输标签和业务标签的形成，以及在边缘节点和中继节点上的流映射配置来处理 DetNet 域内业务。

第3层执行层包括4大功能：“门控操作”模块根据带宽和业务路径，规划对应队列的门操作序列，并调用 SB 插件模块向网络设备下发门操作列表，确保业务的带宽和有边界时延；“带宽调整”模块用于确定性业务的可用带宽管理和实际带宽预留工作；“域间流映射”根据端到端业务整合模块指令，对 Detnet/TSN 域间业务进行封装、流映射规则生成；“队列调整”根据端到端业务整合模块指令，生成各节点对应出端口队列过滤、入队模板生成。

确定性网络系统要想实现“确定性”的 QoS 目标，最终还须依赖转发面设备的芯片能力。

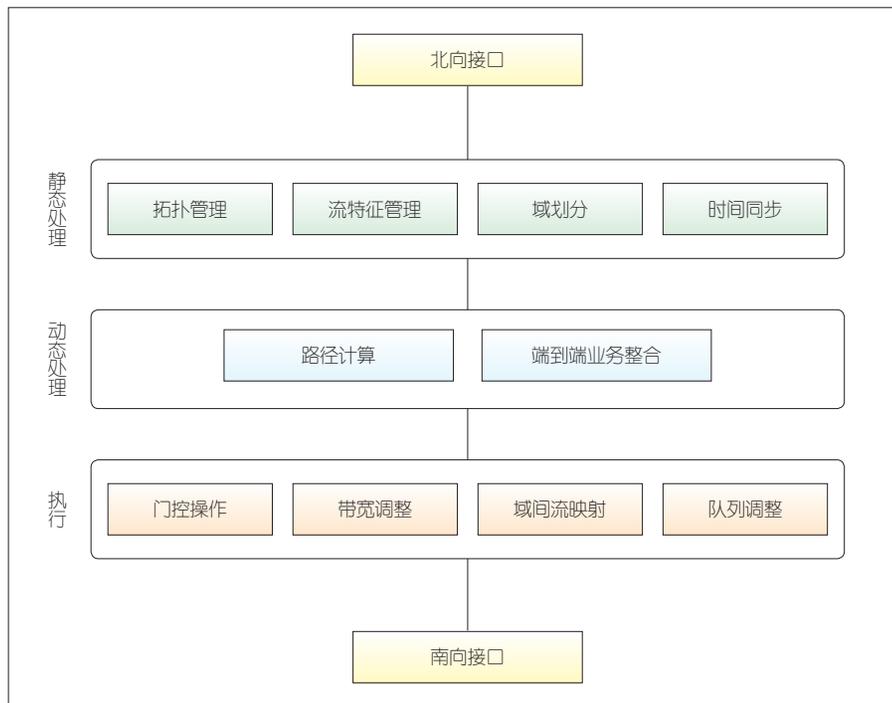
4 结束语

确定性网络应用场景的共性需求，是低且可控的端到端时延保证、可控的时延抖动、极低的丢包率，以及网络规模的动态伸缩。在此基础上，端到端分发协议、标准化的转发行为、

软件定义网络（SDN）集中管理技术、标准化的流信息模型、多时钟域的不同步等还要继续发展，才可能被应用在大规模网络中。多种多样的 OT 网络，会逐步采用 IT 网络所用的通用及泛在网络技术，从而实现 OT 和 IT 基础架构的互通和集成，以提升网络管理效率和路由灵活性，降低技术壁垒和网络成本。

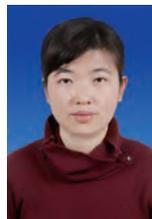
参考文献

- [1] IETF. Deterministic networking architecture: RFC 8655 [S]. 2017
- [2] IEEE. IEEE standard for local and metropolitan area networks, bridges and bridged networks: IEEE 802.1Q [S]. 2018
- [3] IEEE. Specification and management parameters for interspersing express traffic: IEEE 802.3Br [S]. 2016
- [4] 中国信息通信研究院. 车联网白皮书 [R]. 2017



▲图5 确定性网络控制器软件模块

作者简介



魏月华，中兴通讯股份有限公司承载网标准预研总工；从事以太网、IP 路由、云计算数据中心网络、SDN 等技术和标准研究；拥有 15 年以上数据网络产品研发、设计及新技术预研经验；发表论文 3 篇，获授权专利 40 余项。



喻敬海，中兴通讯股份有限公司算法标准部技术预研项目经理；从事确定性业务承载技术、以太网、IP 路由、云计算数据中心网络、SDN 等技术和标准研究；拥有 20 年以上数据网络产品研发、设计及新技术预研经验；发表论文 6 篇，申请发明专利 100 余项。



罗鉴，中兴通讯股份有限公司有线研究院总工；从事数据产品研发、系统设计、网络架构和标准研究以及技术规划等工作。

《中兴通讯技术》杂志（双月刊）投稿须知

一、杂志定位

《中兴通讯技术》杂志为通信技术类学术期刊。通过介绍、探讨通信热点技术，以展现通信技术最新发展动态，并促进产学研合作，发掘和培养优秀人才，为振兴民族通信产业做贡献。

二、稿件基本要求

1. 投稿约定

- (1) 作者需登录《中兴通讯技术》投稿平台：tech.zte.com.cn/submission，并上传稿件。第一次投稿需完成新用户注册。
- (2) 编辑部将按照审稿流程聘请专家审稿，并根据审稿意见，公平、公正地录用稿件。审稿过程需要1个月左右。

2. 内容和格式要求

- (1) 稿件须具有创新性、学术性、规范性和可读性。
- (2) 稿件需采用 WORD 文档格式。
- (3) 稿件篇幅一般不超过 6 000 字（包括文、图），内容包括：中、英文题名，作者姓名及汉语拼音，作者中、英文单位，中文摘要、关键词（3 ~ 8 个），英文摘要、关键词，正文，参考文献，作者简介。
- (4) 中文题名一般不超过 20 个汉字，中、英文题名含义应一致。
- (5) 摘要尽量写成报道性摘要，包括研究的目的、方法、结果 / 结论，以 150 ~ 200 字为宜。摘要应具有独立性和自明性。中英文摘要应一致。
- (6) 文稿中的量和单位应符合国家标准。外文字母的正斜体、大小写等须写清楚，上下角的字母、数据和符号的位置皆应明显区别。
- (7) 图、表力求少而精（以 8 幅为上限），应随文出现，切忌与文字重复。图、表应保持自明性，图中缩略词和英文均要在图中加中文解释。表应采用三线表，表中缩略词和英文均要在表内加中文解释。
- (8) 所有文献必须在正文中引用，文献序号按其在文中出现的先后次序编排。常用参考文献的书写格式为：
 - 期刊 [序号] 作者 . 题名 [J]. 刊名, 出版年, 卷号 (期号): 引文页码 . 数字对象唯一标识符
 - 书籍 [序号] 作者 . 书名 [M]. 出版地: 出版者, 出版年: 引文页码 . 数字对象唯一标识符
 - 论文集析出文献 [序号] 作者 . 题名 [C] // 论文集编者. 论文集名 (会议名). 出版地: 出版者, 出版年 (开会年): 引文页码 . 数字对象唯一标识符
 - 学位论文 [序号] 作者 . 题名 [D]. 学位授予单位所在城市名: 学位授予单位, 授予年份 . 数字对象唯一标识符
 - 专利 [序号] 专利所有者 . 专利题名: 专利号 [P]. 出版日期 . 数字对象唯一标识符
 - 国际、国家标准 [序号] 标准名称: 标准编号 [S]. 出版地: 出版者, 出版年 . 数字对象唯一标识符
- (9) 作者超过 3 人时，可以感谢形式在文中提及。作者简介包括：姓名、工作单位、职务或职称、学历、毕业于何校、现从事的工作、专业特长、科研成果、已发表的论文数量等。
- (10) 提供正面、免冠、彩色标准照片一张，最好采用 JPG 格式（文件大小超过 100 kB）。
- (11) 应标注出研究课题的资助基金或资助项目名称及编号。
- (12) 提供联系方式，如：通讯地址、电话（含手机）、Email 等。

3. 其他事项

- (1) 请勿一稿多投。凡在 2 个月（自来稿之日算起）以内未接到录用通知者，可致电编辑部询问。
- (2) 为了促进信息传播，加强学术交流，在论文发表后，本刊享有文章的转摘权（包括英文版、电子版、网络版）。作者获得的稿费包括转摘酬金。如作者不同意转摘，请在投稿时说明。
- (3) 编辑部地址：安徽省合肥市金寨路 329 号凯旋大厦 1201 室，邮政编码：230061。
- (4) 联系电话：0551-65533356，联系邮箱：magazine@zte.com.cn。
- (5) 本刊只接受在线投稿，欢迎访问本刊投稿平台：tech.zte.com.cn/submission。

中兴通讯技术

(ZHONGXING TONGXUN JISHU)

办刊宗旨:

以人为本，荟萃通信技术领域精英
迎接挑战，把握世界通信技术动态
立即行动，求解通信发展疑难课题
励精图治，促进民族信息产业崛起

双月刊 1995 年创刊 总第 153 期
2020 年 8 月 第 26 卷 第 4 期

主管：安徽出版集团有限责任公司
主办：时代出版传媒股份有限公司
深圳航天广宇工业有限公司
出版：安徽科学技术出版社
编辑、发行：中兴通讯技术杂志社

总编辑：王喜瑜
主编：蒋贤骏
执行主编：黄新明
责任编辑：徐烨
编辑：杨广西、卢丹、朱莉、任溪溪
设计排版：徐莹
发行：王萍萍
外联：卢丹
编务：王坤

《中兴通讯技术》编辑部
地址：合肥市金寨路 329 号凯旋大厦 1201 室
邮编：230061
网址：tech.zte.com.cn
投稿平台：tech.zte.com.cn/submission
电子信箱：magazine@zte.com.cn
电话：(0551)65533356

传真：(0551)65850139
发行范围：公开发行
印刷：合肥添彩包装有限公司
出版日期：2020 年 8 月 10 日
中国标准连续出版物号：ISSN 1009-6868
CN 34-1228/TN
定价：每册 20.00 元