第五届编辑委员会

《电信科学》杂志持证记者：吴娜达

国家广播电视总局

举报电话：(010)83138953

读者热线：(010)81055459/5476

(010)81055598(发行部)

E-mail：dxkx@ptpress.com.cn

《电信科学》微信订阅号

电 信 科 学

DIANXIN KEXUE

(月刊,1956 年创刊)

2018 年(第 34 卷)第 8 期

主 管 单 位　中国科学技术协会

主 办 单 位　中国通信学会

人民邮电出版社

出 版 单 位　北京信通传媒有限责任公司

编 辑 单 位　《电信科学》编辑部

北京市丰台区成寿寺路 11 号

邮电出版大厦 8 层(100078)

主　　编　韦 乐 平

编辑部主任　吴 娜 达

执 行 主 任　李 彩 珊

传　　真　(010)81055494

电　　话　(010)81055443/5459/5476(编辑部)

(010)81055476(广告部)

投 稿 网 址　www.telecomsci.com

正 文 排 版　《电信科学》杂志排版室

印　　刷　北京时捷印刷有限公司

国 内 发 行　北京报刊发行局

订　购　处　全国各地邮局

邮 发 代 号　2-397

国 外 发 行　中国国际图书贸易集团有限公司

国 外 代 号　M841

刊　　号　ISSN 1000-0801 / CN 11-2103/TN

广告经营许可证　京东工商广字第 8032 号

出 版 日 期　2018 年 8 月 20 日

定　　价　68.00 元

目 录

专题:5G

研究与开发

电信科学

第 34 卷第 8 期 2018 年 8 月

Telecommunications Science

August 2018 (Vol.34 No.8)

Chief Editor: WEI Leping
Director: WU Nada
Sponsor: China Institute of Communications
Posts & Telecom Press
Editor: *Telecommunications Science* Magazine Office (F8, You Dian Publisher Building, No.11 Chengshousi Road, Fengtai District, Beijing 100078, China)
General Distribution Office: Beijing Newspapers and Periodicals Distribution Office
Overseas Distributor: China International Book Trading Corporation (P.O.Box 399, Beijing, China)

专题：5G

专 题 导 读

全球信息通信业正迎来新一轮创新浪潮，5G 开启万物互联新时代，将在大幅提升以人为中心的移动互联网业务使用体验的同时，全面支持以物为中心的物联网业务，实现人与人、人与物和物与物的智能互联。

为了及时分析和探讨 5G 和后 5G 技术演进、标准制定、理论攻关、应用开发、测试进程等内容，在 2017 年组织策划的 5G 专题基础上，邀请高等院校、政府监管部门、研究院、运营商等在本领域从事一线研究的权威专家和老师撰文介绍最新的研究成果，以指导 5G 和后 5G 研发方向和理论技术攻关。

5G 标准化工作最近取得了突破性的进展，我国已启动 5G 移动通信行业标准的制定，正通过技术试验促进产业发展，运营商 5G 网络部署的策略也日益清晰。《5G 移动通信技术标准综述》介绍了 5G 标准的主要特点和技术内容，包含新空口、核心网和安全，同时展望 5G 标准发展趋势。

为了满足巨流量、大链接、超低时延等 5G 组网性能需求，针对广覆盖和高容量设计的传统无线接入网络架构亟需演进，《5G 移动通信系统的接入网络架构》阐明了 5G 接入网络架构的特点和重要性，从学术界和产业界两个角度详细介绍了 5G 接入网络架构的设计原理和具体组成，分析了优点和不足，并探讨了接入网络架构的挑战和未来的可能发展方向。

《面向通信与计算融合的 5G 移动增强/虚拟现实》对移动 AR/VR 多级计算模型、智能传输机制和服务时延保障的研究现状和发展趋势进行综述，总结了存在的问题，并提出下一步的研究方向。

《5G 先进技术研究进展》介绍了 5G 的一些热点技术，如大规模天线技术、超密集组网技术和非正交多址技术。

《基于人工智能的无线传输技术最新研究进展》介绍了基于人工智能的信道估计、信号检测、信道状态信息反馈与重建、信道解码、端到端的无线通信技术的最新研究进展，并对利用人工智能的无线传输技术发展趋势进行了展望。

3GPP 第一版 5G 标准已经冻结，面向商用的 5G 网络规模试验工作需要及时启动，以验证 5G 关键技术与性能。《面向商用的 5G 网络关键问题研究及验证》介绍了 5G 规模试验基本情况，对面向商用的 5G 网络关键问题进行了梳理，并提出了验证思路。

5G eMBB 场景控制信道采用 Polar 码，《面向 5G 新空口技术的 Polar 码标准化研究进展》对 Polar 码的基本概念及原理进行了概述，并对近年来国内外研究机构针对 Polar 码开展的标准化研究工作进行了总结分析。

本专题凝聚了作者及其单位在 5G 和后 5G 方面的最新研究成果和心血汗水，希望本专辑可以引起业界同行共鸣，激发我国移动通信从业人员和科技工作者对其中的核心理论和关键技术进行突破创新，从而实现我国在 5G、后 5G 技术和产业的引领。在此，也对各位作者的积极支持和辛勤工作表示衷心的感谢。

专题策划人：彭木根，男，北京邮电大学网络技术研究院副院长、教授、博士生导师，IET 会士，中国通信学会青年工作委员会主任委员，中国电子学会青年科学家俱乐部副主席，北京科技人才研究会副理事长，《电信科学》期刊编辑委员会副主任委员。发表本领域顶级 IEEE 期刊论文 80 余篇，其中 16 篇论文入选 ESI 高被引论文数据库，Google 学术引用 6 600 余次。出版专著译著 12 部，包括英文专著 1 部，获得中华优秀出版物奖图书奖。曾获得北京市科学技术奖一等奖（排名第一）、高等学校科学研究优秀成果奖（科学技术）技术发明奖一等奖（排名第二）和自然科学奖二等奖（排名第一）、中国通信学会技术发明奖一等奖（排名第一）、茅以升科技奖北京青年科技奖、国际电气和电子工程师协会赫兹最佳论文奖和亚太区杰出青年科学家奖等。

专题：5G

5G 移动通信技术标准综述

杜滢，朱浩，杨红梅，王志勤，徐杨
（中国信息通信研究院，北京 100083）

摘　要：业界齐心协力打造能满足移动宽带业务和物联网业务的 5G 技术标准，近期国际标准组织 3GPP 宣布冻结第一个独立组网 5G 标准。5G 具有大带宽、低时延、灵活配置的特点，设计全新的基于服务化系统架构，并具备网络切片、边缘计算等重要业务能力。结合 5G 系统特点，分析了新空口、新核心网和安全机制等内容，同时展望 5G 标准发展趋势。

关键词：新空口；服务化架构；安全
中图分类号：TN929
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018231

Review of 5G mobile communication technology standard

DU Ying, ZHU Hao, YANG Hongmei, WANG Zhiqin, XU Yang
China Academy of Information and Communications Technology, Beijing 100083, China

Abstract: The industry is working together to create 5G technology standard, which should meet the needs of mobile broadband and internet of things. Recently, 3GPP declares the first 5G standard which supports the standalone deployment frozen. 5G supports wider bandwidth, low latency, flexible configuration. New service-based architecture is designed for 5G, which supports important service capabilities, e.g. network slice, edge computing. 5G standards were introduced, including new radio, new core and security. In the end, the trends of 5G standard were summarized.

Key words: new radio, service-based architecture, security

1　引言

5G 是近几年通信产业的研发重点。2012 年全球主要国家和区域纷纷启动 5G 移动通信技术需求和技术研究工作。同期国际电信联盟（ITU）启动了一系列 5G 工作，如 5G 愿景、需求、评估方法等，并于 2015 年 6 月正式发布了 5G 愿景，明确面向 2020 年及未来的移动通信市场、用户、业务应用的发展趋势，并提出未来移动通信系统的框架和关键能力。

5G 开启万物互联新时代。业界一般认为移动通信 10 年一代，2G 时代提供语音和低速数据业务，3G 时代在提供语音业务的同时，开始提供基础的移动多媒体业务，4G 时代提供移动宽带业务，到了 5G 时代，移动通信将在大幅提升以人为中心的移动互联网业务使用体验的同时，全面支

收稿日期：2018−07−03；修回日期：2018−08−10

持以物为中心的物联网业务，实现人与人、人与物和物与物的智能互联[1]。5G 满足增强移动宽带、海量机器类通信和超高可靠低时延通信三大类应用场景，在 5G 系统设计时需要充分考虑不同场景和业务的差异化需求。

新业务新需求对 5G 系统提出新挑战。ITU 定义了八大关键技术指标[2]，其中峰值速率、移动性、时延和频谱效率是传统的移动宽带关键技术指标，新定义了 4 个关键指标，即用户体验速率、连接数密度、流量密度和能效。5G 将满足 20 Gbit/s 的光纤般接入速率、毫秒级时延的业务体验、千亿设备的连接能力、超高流量密度和连接数密度及百倍网络能效提升等极致指标，一个系统如何同时满足多样业务需求，5G 系统设计面临新的挑战。

2　5G 标准规划

3GPP 于 2018 年 6 月发布第一个独立组网 5G 标准。3GPP 制定 R15 和 R16 标准满足 ITU IMT-2020 全部需求，其中 R15 为 5G 基础版本，重点支持增强移动宽带业务和基础的低时延高可靠业务，R16 为 5G 增强版本，将支持更多物联网业务。考虑到 5G 将与 LTE 较长时间共存，并且运营商拥有的频谱不同、部署节奏不同、5G 网络业务定位不同，3GPP 标准分阶段支持多种 5G 组网架构。具体地，R15 包含 3 个子阶段，第一个子阶段为 2017 年年底完成非独立组网的 5G 标准，第二个子阶段为 2018 年 6 月完成可独立组网的 5G 标准，第三个子阶段为 2018 年 12 月完成支持更多组网架构的版本，这些子版本将为运营商提供更多组网选择。2017 年 12 月，3GPP 发布了 R15 非独立组网的标准、5G 核心网架构和业务流程标准，重点增强支持移动宽带业务，5G 基站与 4G 基站或 4G 核心网连接，用户通过 4G 基站接入网络后，5G 新空口和 4G 空口为其提供数据服务，4G 负责移动性管理等控制功能。2018 年 6 月 3GPP 发布了第一个 5G 独立组网标准，5G 基站直接连接 5G 核心网，支持增强移动宽带和基础低时延高可靠业务，基于全服务化架构的 5G 核心网，5G 系统能提供网络切片、边缘计算等新应用。即将于 2018 年 12 月发布的 R15 第三个子阶段标准将完成更多组网架构，支持 4G 基站接入 5G 核心网，以快速提供网络切片、边缘计算等业务能力，之前两个子版本在一定程度上对性能、部署周期和成本等进行折中。此外，3GPP 将于 2019 年年底发布 R16 标准，R16 标准在 R15 的基础上，进一步增强网络支持移动宽带的能力和效率，同时扩展支持更多物联网场景。

3　5G 新空口技术

3.1　物理层和底层协议

5G 新空口（new radio，NR）具有大带宽、低时延、灵活配置的特点，满足多样业务需求，同时易于扩展支持新业务。下面分别介绍 5G 新空口的特点和关键技术[3-6]。

在波形和多址方面，NR 仍采用正交频分多址（OFMA）作为上行和下行基础多址方案，考虑到上行覆盖问题，上行还支持单载波方案 DFT-S-OFDMA，此时，仅支持单流传输。相比于 LTE 系统 90%的频谱利用率，NR 支持更高的频谱利用、更陡的频谱模板，并通过基于实现的新波形方案避免频带之间的干扰。

NR 支持更大带宽。针对 6 GHz 以下的频谱，5G 新空口支持最大 100 MHz 的基础带宽；针对 20~50 GHz 频谱，5G 新空口支持最大 400 MHz 的基础带宽，相对于 LTE 最大 20 MHz 的基础带宽，5G 能更有效地利用频谱资源，支持增强移动宽带业务。此外，5G 新空口采用部分带宽设计，灵活支持多种终端带宽，以支持非连续载波，降低终端功耗，适应多种业务需求。

NR 支持灵活参数集，以满足多样带宽需求。NR 以 15 kHz 子载波间隔为基础，可根据 15×2^{u}

灵活扩展，其中 u=0,1,2,3,4，也就是说 NR 支持 15 kHz、30 kHz、60 kHz、120 kHz、240 kHz 5 种子载波间隔，其中子载波 15 kHz、30 kHz、60 kHz 适用于低于 6 GHz 的频谱，子载波 60 kHz、120 kHz、240 kHz 适用于高于 6 GHz 的频谱。新空口定义子帧长度固定为 1 ms，每个时隙固定包含 14 个符号，因而对于不同子载波间隔，每个时隙长度不同，分别为 1 ms、0.5 ms、0.25 ms、0.125 ms 和 0.062 5 ms。

NR 支持灵活帧结构，定义大量时隙格式，满足各种时延需求。LTE 定义了 7 种帧结构、11 种特殊子帧格式，NR 定义了 56 种时隙格式，并可以基于符号灵活定义帧结构。LTE 帧结构以准静态配置为主，高层配置了某种帧结构后，网络在一段时间内采用该帧结构，帧结构周期为 5 ms 和 10 ms，在特定场景下，也可以支持物理层的快速帧结构调整；NR 从一开始设计就支持准静态配置和快速配置，支持更多周期配置，如 0.5 ms、0.625 ms、1 ms、1.25 ms、2 ms、2.5 ms、5 ms、10 ms，此外，时隙中的符号可以配置上行、下行或灵活符号，其中灵活符号可以通过物理层信令配置为下行或上行符号，以灵活支持突发业务。

NR 支持更大数据分组的有效传输和接收，提升控制信道性能。增强移动宽带业务的大数据分组对编码方案的编译码的复杂度和处理时延提出了挑战，LPDC 在处理大数据分组和高码率方面有性能优势，成为 NR 的数据信道编码方案。对于控制信道，顽健性是最重要的技术指标，极化码 Polar 在短数据分组方面有更好的表现，成为 NR 的控制信道编码方案。

NR 支持基于波束的系统设计，提供更灵活的网络部署手段。LTE 中同步、接入采用广播传输模式，数据信道支持波束成形传输模式。为了实现同步、接入和数据传输 3 个阶段的匹配，NR 中同步、接入、控制信道、数据信道均基于波束传输，并支持基于波束的测量和移动性管理，以同步为例，NR 支持多个同步信号块，SSB 可以指向不同的区域，比如楼宇的高层、中层和地面，为网络规划提供更多可调手段。

NR 支持数字和混合波束成形。低频 NR 主要采用传统的数字波束成形，针对高频 NR，既需要补偿路损，又需要合理的天线成本，因而 NR 引入模拟+数字的混合波束成形。NR 下行支持最大 32 端口的天线配置，上行支持最大 4 端口的天线配置；在具体 MIMO 传输能力方面，下行单用户最大支持 8 流，最大支持 12 个正交多用户，上行单用户最大支持 4 流。另，与 LTE 定义了多种传输模式不同，NR 目前定义了一种传输模式，即基于专用导频的预编码传输模式。此外，相比于 LTE，5G 新空口定义更多导频格式（如 front-loaded 和支持高速移动的额外 DMRS），以支持更多天线阵列模式和部署场景。

NR 实现传输资源和传输时间的灵活可配。支持多种资源块颗粒度，如基于时隙、部分时隙、多个时隙的力度，以满足不同业务需求。支持可配置的新数据分组传输和重传时序，在满足灵活帧结构的同时，满足低时延需求。

预计 5G 新空口预计将部署在较高频段，考虑到基站和终端天线配置的差异，需要重点研究如何保障 5G 上行覆盖。最直接的方案是提升终端发射功率，此外，考虑到 5G NR 将在较长时间内和 LTE 共存，可以利用低频 LTE 上行资源保障系统上行覆盖。主要有两类方案：在 5G 业务信道覆盖受限的情况下，回退到低频 LTE 业务信道来保证上行覆盖，如双连接或切换；在 5G 业务信道覆盖受限的情况下，通过补充上行（SUL）保证上行覆盖，即占用部分低频 LTE 上行资源传输 NR。

3.2 高层协议

NR 高层协议大量重用了 LTE 设计。下面分为控制面和用户面分别介绍 NR 高层协议。

控制面，NR 与 LTE 有三大主要差异。相比于 LTE，NR 新增了 RRC inactive 状态，该状态下，终端、基站和核心网部分保留 RRC 和 NAS 上下文，这样可以快速进入 connected 状态，在省电的同时，降低连接时延、减少信令开销和功耗，以适应未来各种物联网场景。在 LTE 和 NR 双连接架构中，扩展了 NR 的 RRC 协议，新增支持 RRC 分集模式，即辅小区复制主小区的 RRC 信息，并通过主小区和辅小区同时向终端发送 RRC 信息，从而提升手机接收 RRC 消息的成功率和可靠性。此外，LTE 仅支持广播发送系统信息，NR 系统信息支持基于请求和广播两种方式，以降低网络广播开销，并提升系统前向兼容性，扩展资源承载类型。

相比于 LTE，NR 增强协议栈功能和性能。NR 支持 6 种承载类型，以提升接入网的组网灵活度。为了提高数据可靠性，5G 核心网支持基于 IP 流的 QoS 控制，实现更灵活和更精细的 QoS 控制，为了实现端到端 QoS，NR 新增 SDAP 层，执行 IP 流和无线承载间映射。

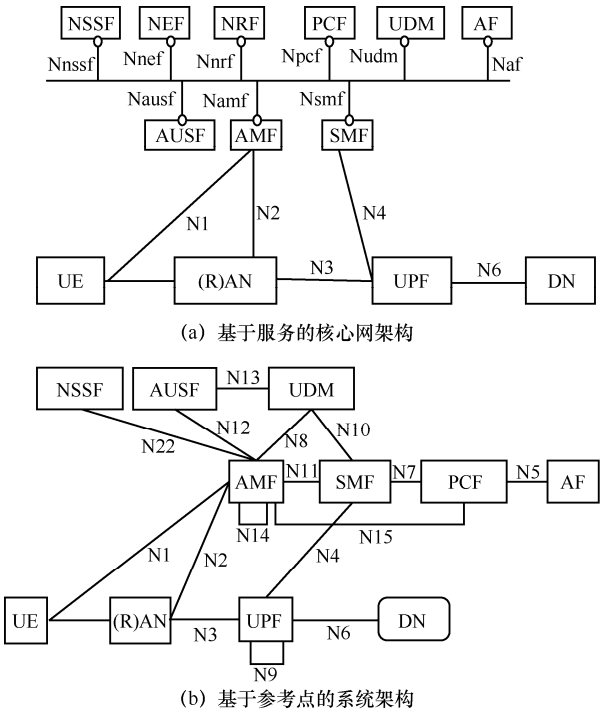
此外，NR 提供更灵活的接入网架构。除了支持与 LTE 相同的接入网架构，5G 支持中心单元/分布单元（CU/DU）分离的接入网架构，其中 CU 为集中控制，DU 为灵活部署。

4　新核心网技术

5G 核心网标准包括新的总体架构和协议模型，针对移动宽带数据服务提供优化的用户接入、会话管理、服务质量、策略控制以及应用与网络交互能力等基础网络内容，还标准化了端到端网络切片、靠近无线网的边缘计算应用。

4.1　5G 核心网架构与接口

为支持差异化的 5G 应用场景和云化部署方式，5G 采用全新的基于服务化系统架构。系统架构中的元素被定义为一些由服务组成的网络功能，这些功能可以被部署在任何合适的地方，通过统一框架的接口为任何许可的网络功能提供服务。这种架构模式采用模块化、可重用性和自包含原则来构建网络功能，使得运营商部署网络时能充分利用最新的虚拟化和软件技术，以细粒度的方式更新网络的任一服务组件，或将不同的服务组件聚合起来构建服务切片。图 1（a）显示了服务化架构的设计原则，同时 Stage 2 规范还提供了基于参考点的系统架构（图 1（b）），其更注重描述实现系统功能时网络功能间的交互关系。



(a) 基于服务的核心网架构

(b) 基于参考点的系统架构

图 1　5G 核心网系统架构

4.2　5G 核心网功能

如图 1 所示，5G 核心网主要的网络功能如下。

（1）接入控制和移动性管理功能（AMF）

主要提供网络接入控制、接入和移动性管理等功能，是 NAS 信令的终结点。5G AMF 针对不同类型的用户终端提供终端能力参数、不同的移动性策略和模式，并以此为依据提供优化的连接管理和寻呼优化。

（2）会话管理功能（SMF）

核心网使用 PDU 会话来标识终端到某个数据

网络间的数据业务连接，5G 支持的 PDU 会话类型包括 IPv4、IPv6、以太网和无结构。

（3）网络切片选择功能（NSSF）

根据用户签约和 UE 上报的候选切片选择辅助信息 NSSAI 来为 UE 选择一个服务切片实例，并为 UE 指派提供服务的 AMF 集合。

（4）策略控制与计费功能（PCF）

在 5G 系统进行了扩展，从 4G 单纯地针对业务数据流，扩展到覆盖用户接入移动性以及终端选路的策略控制。5G 系统架构的 QoS 模型细化到每一个五元组的粒度，且由用户面标签直接实现，无需额外信令，使不同的数据服务能够有效利用无线资源，以支持各种应用需求。

（5）统一数据库功能（UDR）

5G 系统架构引入统一的结构化（UDR）和非结构化数据库（UDSF）功能，将业务逻辑处理和数据存储分离，网络切片的服务弹性更高，负载聚合和容灾机制更灵活。

（6）用户面功能（UPF）

采用控制和用户面分离的模型，由 SMF 进行管理，实现灵活的业务流路径编排和有服务质量保证的转发。

（7）网络功能库（NRF）

提供服务的注册、管理和查询功能。

4.3 5G 核心网业务

相比于 4G 网络通过“专有核心网”的特性支持网络切片，5G 网络切片是一个更强大的概念。在 3GPP 5G 系统架构的范围内，网络切片是指一组 3GPP 定义的特征和功能，它们组成向 UE 提供服务的一个完整 PLMN。网络切片使得网络运营商能够在统一的云化基础设施上部署多个独立的 PLMN，其中每个网络切片只需实例化所属签约用户所关注的特性、功能和业务。

图 2 展示了 3GPP 网络切片的更多细节。在图 2 中，网络切片#3 是直接部署，其中所有网络功能仅服务于单个网络切片。还展示了一个 UE 如何从多个网络切片＃1 和＃2 获得服务。在这样的部署中，一组切片可以共享一些网络功能，包括 AMF、相关的策略控制（PCF）以及网络功能库（NRF）。用户面业务，特别是数据业务，可以通过多个独立的网络切片获得。切片＃1 向 UE 提供访问数据网络＃1 的服务，切片＃2 向 UE 提供访问数据网络＃2 的服务。除了用户所有业务共同使用的接入和移动性控制（AMF）的交互之外，这些切片和数据业务彼此独立。可以为每个切片定制例如不同的 QoS 数据业务或不同的应用功能，全部通过策略控制框架来确定。

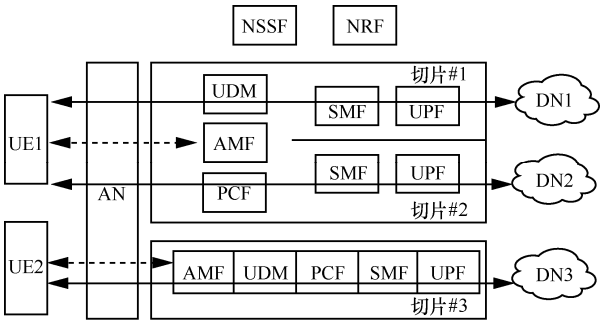


图 2 5G 核心网切片架构

3GPP 边缘计算支持应用功能在网络拓扑灵活的按需部署，以优化时延和传输网络负载。主要优化功能包括会话和业务连续性（SSC）模式或用户面的上行分类器、分支点。SSC 模式包括传统的模式（SSC 1）和新模式，SSC 1 是在 UE 位置变化时仍然维持 IP 地址矛盾稳定来持续支持应用并维护到 UE 的路径。新的模式允许 IP 地址矛盾的重置，包含两种方式：先建后断（SSC mode 3）和先断后建（SSC mode 2）。这种架构使得应用可以影响合适的数据业务特性和 SSC 模式的选择。

由于 5G 网络部署预计将服务于海量的移动数据流量，因此高效的用户面路径管理至关重要。除了 SSC 模式之外，系统架构还定义了上行链路分类器和分支点的功能，以允许在 IP 锚点之前在用户面路径上有选择性地卸载和插入数据流。而且在策略允许时，应用功能可提供优化数据流路

由相关的信息实现和网络协调，或者向 5G 系统订阅可能与应用相关的事件。

R15 版本的 5G 系统将全面采用基于 IMS 的分组语音方案，考虑到 5G 形成满足语音连续性要求的覆盖能力需要一个过程，5G 语音方案需要考虑使用户尽可能驻留 5G 网络监听语音呼叫，并根据网络质量来选择建立语音业务的路径。总的来说，5G 提供 VoNR 和 EPS fallback 两种语音机制。

（1）VoNR

用户驻留在 5G 小区，接入 5G 核心网完成 IMS 语音业务注册，当要发起入呼/出呼业务时，终端在 NR 基站和 UPF 间建立 QCI=1 的语音专用业务流，实现语音接续。呼叫过程中如果移出 5G 覆盖，则进行 PS 切换操作，到 LTE 网络重建语音和数据业务会话。

（2）EPS fallback

用户驻留在 5G 小区，接入 5G 核心网完成 IMS 语音业务注册，当要发起入呼/出呼业务时，基站侧将拒绝在 5G 系统建立语音业务流的请求，同时触发 PS 切换或重定向过程，到 4G 网络完成语音业务承载的建立。由于 PGW 与 UPF 是合设的，所以呼叫过程可以保持 IP 地址和业务连续性。

基于自组织的分布式网络管理模型如图 3 所示。

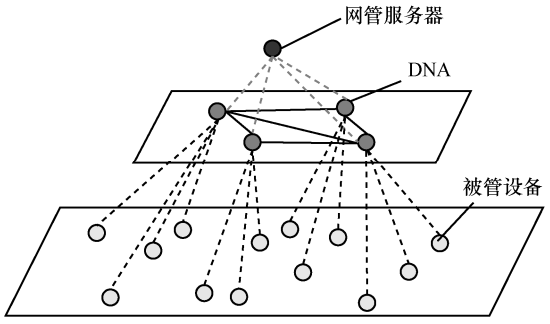


图 3　基于自组织的分布式网络管理模型

5　5G 安全技术

5G 网络新的发展趋势，尤其是 5G 新业务、新架构、新技术，对安全和用户隐私保护都提出了新的挑战。5G 安全机制除了要满足基本通信安全要求之外，还需要为不同业务场景提供差异化的安全服务，能够适应多种网络接入方式及新型网络架构，保护用户隐私，并能提供开放的安全能力。因而 5G 网络安全设计目标如下。

- 提供统一的认证框架，支持多种接入方式和接入凭证，从而保证所有终端设备安全地接入网络。
- 提供按需的安全保护，满足多种应用场景中的终端设备的生命周期、业务的时延要求。
- 提供隐私保护，满足用户隐私保护以及相关法规的要求。

5G 安全除了应保护多种应用场景下的通信安全以外，还应能保护 5G 网络架构本身的安全。5G 网络架构的重要特征包括 NFV/SDN、网络切片以及能力开放。因此，5G 安全应保证 NFV/SDN 引入之后移动网络的安全，NFV/SDN 技术实现了软件与硬件的解耦，NFV 技术的部署使得部分功能网元以虚拟功能网元的形式部署在云化的基础设施上，网络功能由软件实现，不再依赖于专有通信硬件平台，因此，5G 安全需要考虑 5G 基础设施的安全，从而保障 5G 业务在 NFV 环境下能够安全运行。另外，5G 网络中通过引入 SDN 技术提高数据传输效率，实现更好的资源配置，需要考虑 SDN 控制网元和转发节点的安全隔离和管理以及 SDN 流表的安全部署和正确执行；再者，5G 网络通过建立网络切片，为不同业务提供差异化的安全服务，根据业务需求针对切片定制其安全保护机制，实现客户化的安全分级服务，所以 5G 安全还应保证网络切片的安全，包括切片安全隔离、切片的安全管理、UE 接入切片的安全、切片之间通信的安全等；还有，5G 网络的能力开放功能部署于网络控制功能之上，以便网络服务和管理功能向第三方开放，能力开放不仅体现在整个网络能力的开放上，还体现在网络内部网元之间的能力开放上，与 4G 网络的点对点流程定义不

同，5G 网络的各个网元都提供了服务的开放，不同网元之间通过 API（应用程序接口）调用其开放的能力，所以 5G 安全应能保证能力开放的安全，既能保证开放的网络能力安全地提供给第三方，也应能保证网络的安全能力（如加密、认证等）可以开放给第三方使用。

5G 网络安全架构[7]需满足 5G 多样化业务场景和新技术新特征引入的新的安全需求和挑战。5G 网络安全架构的设计原则包括支持数据安全保护、体现统一认证框架和业务认证、满足能力开放以及支持切片安全和应用安全保护机制。5G 网络安全架构如图 4 所示。

图 4 中，将 5G 网络安全架构分为以下 8 个安全域：网络接入安全，保障用户接入网络的数据安全；网络域安全，保障网元之间信令和用户数据的安全交换；首次认证和密钥管理，包括认证和密钥管理的各种机制，体现统一的认证框架；二次认证和密钥管理，UE 与外部数据网络（如业务提供方）之间的业务认证以及相关密钥管理；安全能力开放，体现 5G 网元与外部业务提供方的安全能力开放，包括开放数字身份管理与认证能力。另外，通过安全开放能力，5G 网络也可以获取业务对于数据保护的安全需求，完成按需的用户面保护；应用安全，保证用户和业务提供方之间的安全通信；切片安全，体现切片的安全保护，例如 UE 接入切片的授权安全、切片隔离安全等；安全可视化和可配置，体现用户可以感知安全特性是否被执行，这些安全特性是否可以保障业务的安全使用和提供。

5G 网络安全技术标准确定了 5G 系统安全架构和流程相关要求，主要包括安全框架、接入安全、用户数据的机密性和完整性保护、移动性和会话管理安全、用户身份的隐私保护以及与 EPS（演进的分组系统）的互通等相关内容。5G 安全采用可扩展认证协议（EAP）框架实现统一认证，支持用户在接入网间无缝切换，同时，通过增强的安全机制进行用户隐私保护（如身份标识等），并支持按需的用户数据保护方法。

此外，NFV/SDN 等新技术将会给 5G 网络安全带来新的影响，ETSI NFV 安全组的研究内容涉及 NFV 安全架构、隐私保护、合法监听、MANO（管理和编排）安全、证书管理、安全管理、安全部署等方面；ONF（开放网络基金会）以及 ITU-T 的研究内容涉及 SDN 安全的标准化工作。

6 5G 标准发展趋势

5G 标准将持续发展。5G 第二版本（R16）标准的预研已经启动，R16 在增强基础的移动宽带业务能力和基础网络架构能力的同时，重点提升对垂直行业应用的支持，特别是对低时延高可靠

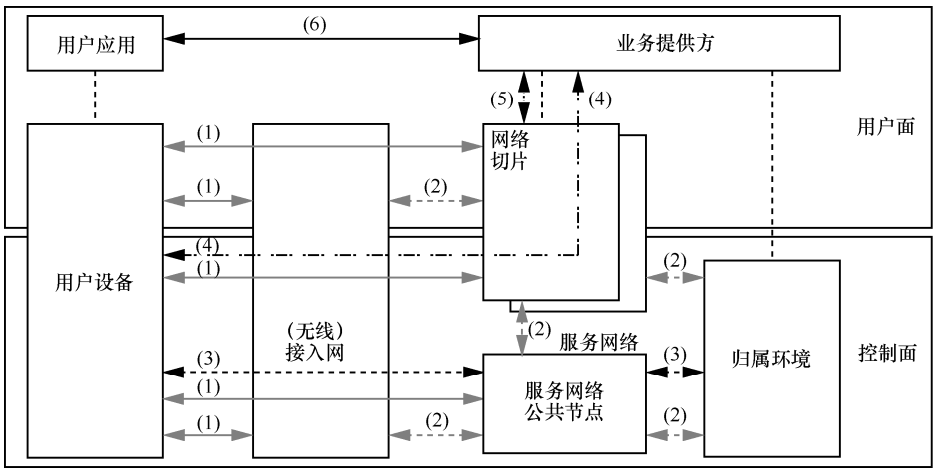


图 4 5G 网络安全架构示意

类业务的支持。在移动宽带业务能力方面，R16 将重点研究多天线增强、载波聚合与大带宽增强、远端干扰删除以及已经开展的非正交多址和免许可 5G 技术。在基础能力提升方面，重点研究终端节能、定位增强、移动性增强、基于 RAN 的大数据收集与应用、服务化架构增强、智能化运营、切片增强等内容。在垂直行业应用方面，研究 5G 车联网（NR V2X）、低时延高可靠（uRLLC）增强、工业物联网增强（NR IIoT）、NB-IoT/mMTC 增强等内容。

7　结束语

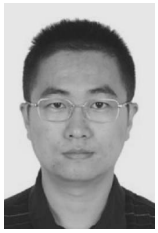
支持万物互联的 5G 标准已经发布，端到端服务化的 5G 网络将更好地服务于人和物的需要。目前，我国正通过技术试验促进产业发展，运营商 5G 网络部署的策略也日益清晰，我国已启动 5G 移动通信行业标准的制定，5G 产业进入快行道。

参考文献：

[1] IMT-2020. 5G 愿景与需求白皮书[R]. 2014.
IMT-2020. 5G vision and demand white paper[R]. 2014.
[2] ITU. 2410-2017-MSW-E-Minimum requirements[S]. 2017.
[3] 3GPP. NR; physical channels and modulation: TS38.211[S]. 2018.
[4] 3GPP. Physical layer procedures for control: TS38.213[S]. 2018.
[5] 3GPP. Physical layer procedures for data: TS38.214[S]. 2018.
[6] 3GPP. Physical layer measurements: TS38.215[S]. 2018.
[7] 3GPP. Security architecture and procedures for 5G system: TS33.501 (v15.1.0)[S]. 2018.

[作者简介]

杜滢（1978−），女，中国信息通信研究院技术与标准研究所副主任，长期从事移动通信无线新技术研究、国际标准研制与仿真评估工作。



朱浩（1982−），男，中国信息通信研究院技术与标准研究所高级工程师，长期从事移动通信网络新技术研究、标准研制和测试工作。

杨红梅（1974−），女，中国信息通信研究院技术与标准研究所主任工程师，长期从事移动通信核心网技术研究、安全技术研究及国际国内标准研究和相关测试工作。

王志勤（1970−），女，中国信息通信研究院副院长，负责移动通信产业和发展策略研究工作。

徐杨（1983−），女，现就职于中国信息通信研究院技术与标准研究所，长期从事移动通信技术研究工作。

专题：5G

5G 移动通信系统的接入网络架构

项弘禹，张欣然，朴竹颖，彭木根
（北京邮电大学，北京 100876）

摘　要：为了满足巨流量、大链接、超低时延等 5G 组网性能需求，针对广覆盖和高容量设计的传统无线接入网络架构亟需演进。首先结合 5G 愿景与需求，阐明了 5G 接入网络架构的特点和重要性；然后从学术界和产业界两个角度详细介绍了 5G 接入网络架构的设计原理和具体组成，分析了优点和不足；最后，探讨了接入网络架构的挑战和未来的可能发展方向。

关键词：5G；网络架构；无线接入网络
中图分类号：TN929.5
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018230

Network architecture in the 5G mobile systems

XIANG Hongyu, ZHANG Xinran, PIAO Zhuying, PENG Mugen
Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: In order to meet the performance requirements of 5G networking such as huge traffic, large links, and ultra-low latency, the traditional wireless access network architecture for wide coverage and high capacity design needs to evolve. Firstly, combining the 5G vision and needs, the characteristics and importance of the 5G access network architecture were clarified. Then the design principle and specific composition of the 5G access network architecture was introduced from the perspectives of academia and industry, and the advantages and disadvantages were analyzed. Finally, the challenges of access network architecture and possible future development directions were discussed.

Key words: 5G, network architecture, radio access network

1　引言

5G 移动网络时代的日益靠近，使得全球产业界和学术界团体加速对于 5G 网络架构的研究，尽早推出 5G 第一个商业版本。3GPP（Third Generation Partnership Project）早在 2016 年就公布了 5G 的两种网络架构[1]：独立组网，即接入网络仅包括新空口（new radio）或 evolved E-UTRA；非独立组网，接入网络中新空口与 evolved E-UTRA 共存。并于 2018 年 6 月完成了 5G NR R15 的第二个版本，同时展开了 R16 版本标准化工作，这极大地提升了业界对于 5G 的信心，对 5G 的后续标准推进和产业发展产生了重大影响。

2　5G 网络愿景与架构特点

2.1　5G 需求与挑战

随着物联网、车联网等技术的蓬勃发展，网络中的接入终端种类与数量不断增长，5G 无线网

收稿日期：2018−06−03；修回日期：2018−08−10

络将实现真正的“万物互联”。预计到 2030 年，全球移动通信设备总数将达到 1 000 亿量级，移动数据流量相较 2010 年增长约 20 000 倍[2]。移动网络中的数据流量呈现分布不均匀，随时间、地点和应用变化多样的特点，为网络传输带来了巨大的压力。另一方面，未来移动通信网络中出现了许多新兴业务，既包含小数据分组服务（如低数据速率的机器通信和实时远程控制），又包含丰富的内容服务（如高清视频、增强现实和在线游戏）。5G 划分出了 3 种业务类型以应对多样化业务服务的差异化性能指标带来的挑战，如下所示[3]。

（1）eMBB（enhanced mobile broadband）

主要包括车站、体育场等超密集区域的巨大数据流量的热点高容量场景。该类场景下性能需求包括 1 Gbit/s 用户体验速率、数十 Gbit/s 峰值速率和数十 Tbit/(s·km^2)的流量密度，网络的流量过载使得现网的流量传输方法面临严峻挑战。此外，eMBB 还包括需要保证用户在高移动性情况下的业务连续性的连续广域覆盖场景，挑战在于随时随地为用户提供 100 Mbit/s 以上的用户体验速率，保证业务的连续性与网络的基本服务能力。

（2）uRLLC（ultra-reliable and low latency communication）

主要面向对时延和可靠性具有极高指标需求的应用，例如车联网、工业控制等低时延高可靠场景，需要网络为用户提供毫秒级的端到端时延和接近 100% 的业务可靠性保证，这与 4G 网络百毫秒级的端到端时延和业务中断时间相距甚远，要求 5G 网络针对更高的可靠性与更低的时延要求提出关键的使能技术。

（3）mMTC（massive machine type communication）

主要应用于机器间通信，以传感器为主，包括智能城市、森林防火和可穿戴设备等低功率大连接场景，满足接入设备数量巨大且功耗极低的需求，预期达到 100 万/km^2 的连接数密度的性能指标。海量的连接数使得网络的控制面负载急剧增加，信令拥塞将是亟待解决的问题。

为了灵活地支撑多样化的业务服务，满足不同应用场景下的性能指标需求，未来移动网络需要具备网络功能和操作管理的多样性，能够智能感知用户需求，对网络功能进行简化、重构和编排，提供高效灵活的网络控制和转发功能，实现不同用户场景、商业模型下各种应用的使用。同时，为方便实现接入网拓扑的部署和维护，5G 网络还需要能够提供按需定制服务，开放网络能力以提供灵活的业务部署环境，在满足差异化业务需求的同时，提升网络服务价值，以实现更友好的网络生态。

除了性能需求带来的挑战，网络效率需求，如频谱效率、能量效率和成本效率也将是 5G 重点关注的效率因素，二者共同定义了 5G 的关键能力。在 5G 中，网络需要实现超百倍的能量效率提升和比特成本降低以及 5~15 倍的频谱效率的提升，以保证 5G 的可持续发展[4]。除了网络性能和效率以外，移动通信网络还面临着感知和开放能力不足的挑战。当前移动通信网络缺乏对用户和业务感知能力，有限的网络开放能力无法实现网络资源与业务需求的友好对接，不利于业务体验的改善和网络运营效率的提升。网络中日益增长的终端设备数量大幅提升了运维成本，降低了运维效率。网络运行过程中会源源不断地产生海量数据，海量数据尚未得到充分利用，造成了数据价值的浪费。5G 通信网络需要充分利用网络运行过程中产生的大量数据降低网络建设成本，提升网络运营水平，提供更加智能的网络运营能力。

2.2　网络架构特征

和以往的网络不同，5G 网络将不只是网络的演进提升，更会带来革命性的改变。5G 网络除了各方面性能的提升，网络架构也将引入许多新的特征。

（1）面向虚拟化网络的 NFV/SDN

NFV（network function virtualization）[5]的引

入将 5G 网络构建成一个虚拟化的网络环境，差异化的软件功能经过虚拟化后运行在相同的硬件设备上，不同网络功能将共享硬件的计算、存储与通信资源。SDN（software-defined networking）[6]的引入则使得网络的可编排性得到提升，分离网络的数据与控制面。参考文献[7]在介绍了欧洲电信标准化协会（ETSI）组织关于 NFV 技术的研究以及开放网络基金会（ONF）组织关于 SDN 技术的研究后，以核心网为例，进一步探讨了二者与 5G 核心网络的有效结合。参考文献[8]则研究了 NFV/SDN 在回传网络上的应用，提出了一种光纤与无线网络融合的 5G-Xhaul 架构，并实现灵活的网络功能分割。

（2）面向多样化服务的网络切片

针对不同的服务需求和性能指标，网络被划分成网络功能实体的逻辑组合，被切片后的网络，即网络切片[9]用于为目标用户和终端提供指定的服务。参考文献[10]在总结现网应对 5G 服务时的不足后，提出一种面向服务的网络用户面编排架构。参考文献[11]则将网络切片的概念进一步提升，加强网络的开放程度，提出“Anything as a Service”，通过快速灵活地调度网络资源，实现服务的动态创建与管理。

（3）面向多维度资源融合的云雾协同

以往的分层异构和云无线接入等网络，受限于集中式的云处理网络架构，性能难以满足 5G 多样化的通信需求。利用边缘节点计算存储和信号处理功能，将雾作为云的协同部分，能够实现集中分布自适应的高效组网。参考文献[12]提出云雾协同架构，将不同服务解析成服务功能链，服务功能链由云雾中的网络功能和多维度资源部署实现。类似地，参考文献[13]提出将网络的多维资源虚拟化，通过灵活分层与编排构成云雾，满足 5G 需求。

3 5G 网络架构研究现状

3.1 5G 网络标准化进展

国际电信联盟（ITU）于 2015 年启动 5G 国际标准制定的准备工作。首先开展 5G 技术性能需求和评估方法研究，明确候选技术的具体性能需求和评估指标；2017 年正式启动 5G 候选技术征集；2018 年底启动 5G 技术评估及标准化；计划在2020年底形成商用能力。在该路线图的指导下，全球范围内的标准化组织和 5G 研究项目组相继开展工作，推动 5G 技术的研发进程。

3GPP 作为 5G 研发的主要标准化组织，自 2015 年 12 月启动 5G 相关议题讨论，制定了 5G 标准化时间表，计划于 2020 年商用。3GPP 5G 的整体研究工作将包含 3 个阶段：研究阶段、工作阶段 1、工作阶段 2，分别对应 R14~R16。2017 年 12 月，3GPP 完成非独立组网的 5G 新空口规范，包含无线接入网络、业务与系统、核心网与终端 3 部分。次年 6 月，完成独立组网的 5G 新空口规范，同时批准 5G 第二阶段新项目，展开 R16 的研究和标准化工作。

5GPPP 作为欧盟框架 7（FP7）项目中的 5G 后续项目，以设计 5G 通信网络和服务为终极目标，于 2015 年 7 月开始展开工作。5GPPP 项目共分为 3 个阶段，第一阶段已于 2017 年年底完成，内容包含新型网络架构等 19 个研究项目；第二阶段包含边缘计算等 21 个研究项目；第三阶段自 2018 年 6 月开始，包含基础设施、自动驾驶、垂直行业测试验证 3 个项目，目前正在积极开展技术评估和测试验证工作。

3.1.1 5G 网络整体架构

2015 年年底，3GPP 系统架构工作组（SA2）正式启动 5G 网络架构的研究课题“NexGen”[14]，明确了 5G 架构的基本功能愿景。2017 年 12 月，3GPP 冻结“NexGen”阶段 2 的工作，输出第一版的 3GPP 5G 网络架构标准[15]，在该标准中，3GPP 定义了 5G 网络整体架构的特点、功能和服务。强调与 4G 系统不同，5G 系统将采用基于服务的架构，并且支持端到端的网络切片功能。

5G 网络架构采用基于服务的架构和通用接

口，传统网元功能基于 NFV 技术拆分成若干个自包含、自管理、可重用的网络功能服务模块，通过灵活定义服务模块集合，实现定制化的网络功能重构，对外通过统一的服务调用接口组成业务流程。图 1 给出了非漫游场景下基于服务的 5G 网络架构示意图。在该架构中，根据特定场景需求，将不同网络功能按需有序组合，实现网络的能力与服务的定制化，为不同业务部署专用网络，实现 5G 网络切片。网络切片技术使运营商能够更加灵活、快速地响应客户需求，支持网络资源的灵活分配。

5GPPP 于 2017 年 12 月发布了“5G 架构 2.0”白皮书[16]，描述了 5G 整体架构设计和评估分析。白皮书进一步地指出，考虑到 5G 需要支持多样化业务和性能指标，网络切片概念需与 5G 系统深度融合。为了支持网络切片功能，5G 架构分为以下功能层：服务层、管理编排层、控制层、多域网络操作系统工具、数据层，分别负责租户应用服务交互和商业策略决策、切换管理与编排、专用网络功能控制、网络资源虚拟化、数据传输功能。各层通过有效的协作完成切片的生命周期管理和整体网络的控制管理等。在所提 5G 架构中，网络切片的开放程度可以分为两种：为租户提供可自主控制和运营的虚拟基础设施，即基础设施作为服务；生成服务实例，直接为租户提供网络服务。

3.1.2　5G 接入网络架构

2017 年 3 月，3GPP 无线接入网络工作组正式开启了 5G NR 工作项目阶段[17]。同年 12 月，完成非独立组网的 5G 新空口规范。2018 年 6 月，完成独立组网的 5G 新空口规范，至此完成了 5G 标准第一阶段的工作，定义了 5G 接入网的整体架构与接入节点架构[18]。

3GPP 定义了新型无线接入网络 NG-RAN，包含两种接入节点：gNB，即提供 5G 控制面和用户面服务的 5G 基站；ng-eNB，为用户提供 LTE/E-UTRAN 服务的基站。gNB 和 ng-eNB 间通过 Xn 接口进行连接，gNB 和 ng-eNB 通过 NG 接口与核心网（5GC）连接。5G gNB 可进一步划分为 CU（central unit）和 DU（distributed unit），提供低成本部署，支持负载管理、实时性能优化在内的协作。NG RAN 的一个显著特点是可以运行独立组网和非独立组网，运营商可根据网络需求和成本灵活选择 5G 部署方式。在独立组网方式下，gNB 连接到 5G 核心网络（5GC）；在非独立组网方式中，利用双连接技术将 NR 和 LTE 紧密集成，连接到现有的 4G 核心网（EPC）。在双连接架构中，主节点和辅助节点同时为用户提供无线资源，提高用户的体验速率。5G 无线协议栈包含两部分：传输用户数据（IP 分组）的用户面和控制信令交互的控制面。用户面引入了服务数据自适应协议层（SDAP），用以支持 5G 核心网基于流的新 QoS 模型。SDAP 层可将带有 QoS 需求的 IP 流映射到特定配置的无线电承载上，在无 RRC 信令辅助的情况下进行动态的配置、重配置。

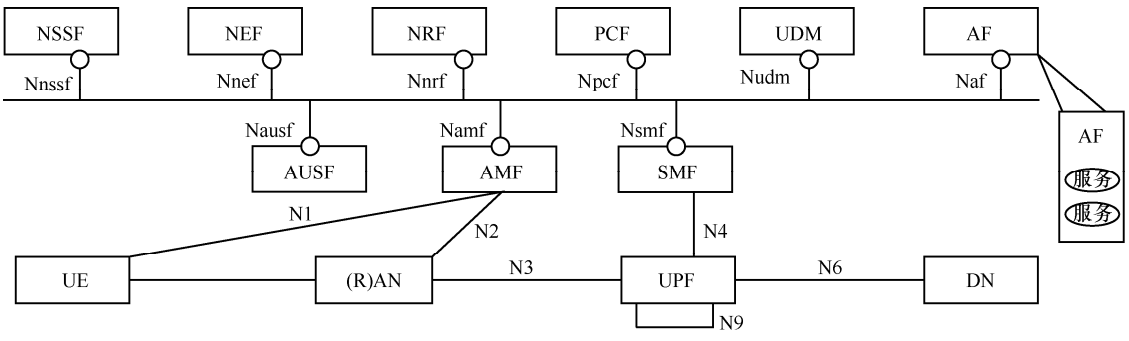


图 1　5G 网络架构与网络功能服务模块化

控制面引入了 RRC inactive 状态，该状态下用户在省电的同时，可更快与 connected 状态切换。

类似地，5GPPP 也对 5G 接入网架构及其关键技术展开了研究[23]。根据部署场景，网络部署将选择合适的用户面和数据面切分方案。包括数据面/控制面完全集中、数据面/控制面部分集中、控制面完全集中、数据面部分集中、数据面/控制面完全分离的 5 种切分选项，其中，完全分离的方案对时延、速率没有要求，各节点通过 X2 接口进行分布式协作。完全集中式方案在多个节点间进行集中式调度，可获得更高的自由度。节点间的协议栈聚合方法将依据网络状况和业务需求而定：无线接入技术（RAT）的聚合点默认选择时间无关的协议栈层（即非时间同步的 RRC/PDCP 层），然而需时间同步的 MAC 层切分方案可实现更高的 RAT 协作。具体地，在 PDCP 层聚合方案中，CU/DU 选择在 PDCP/RLC 层切换，DU 可支持一种和多种空口技术；在 MAC 层聚合方案中，引入了扩展动态频谱访问 MAC 框架，将 MAC 层进一步划分为高/低层 MAC，低层 MAC 对应特定的空口技术。

3.2 网络架构理论研究

面向不同的应用场景与性能需求，无线接入网络需要由传统的以基站为中心，基站—用户式孤立管道传输数据的结构，转变为以用户和业务为中心，异构节点共存与协作的网络结构。5G 网络架构中，有线和无线连接将被有效利用完成回传，提升用户体验速率并降低业务传输时延，满足未来多样化业务的性能需求。

（1）云无线接入网络（C-RAN）

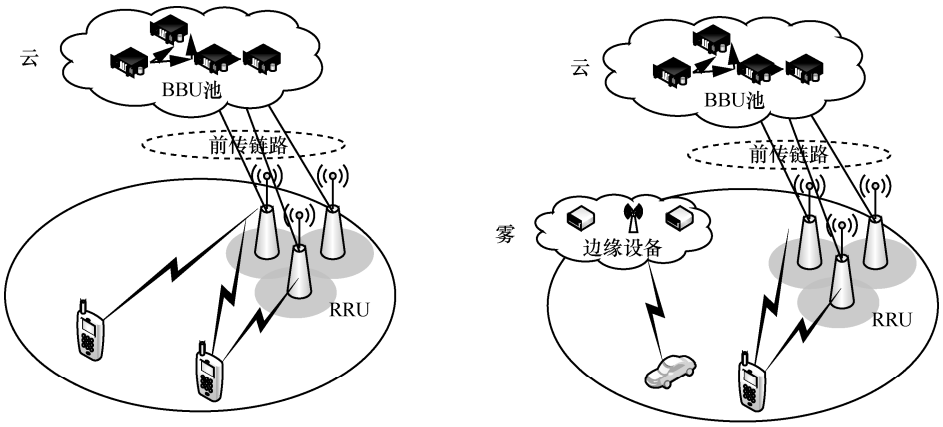
集中式云无线接入网络（cloud RAN，C-RAN）是未来无线接入网演进的重要方向。C-RAN 中，基于通用硬件平台的云计算承载基站的上层基带处理功能，池化的基带处理中心（base band unit pool，BBU 池）提供协作信号处理、集中的资源管理、多 RAT 管理等功能；基站剩余的射频等底层功能被分配给远端无线处理单元（remote radio unit，RRU），汇聚小范围内的 RRU 信号，经部分基带处理后进行前端数据传输。在满足一定的前传和回传网络的条件下，C-RAN 可以有效提升网络容量，降低网络能耗和部署成本。

传统 C-RAN 架构[19]中，BBU 池汇聚了基站大部分功能，RRU 仅负责射频信号处理，这不仅要求网络前传链路能够提供低时延高速率的理想条件，还使得网络架构僵化固定。参考文献[20]提出一种前传资源灵活分配的方法，实现面向业务流量和网络环境的 BBU-RRU 灵活适配。借助 SDN，BBU 池由原有的硬件基础设施变为软件定义的环境，参考文献[21]提出 BBU-RRU 功能可编排，BBU 池不仅能够灵活扩容，还能面向业务的灵活扩展与拆分功能，平衡了实时性和传输网络性能要求。进一步地，参考文献[22]搭建了基于通用处理器的 C-RAN，验证了 C-RAN 在不同环境和传输策略下的性能。

C-RAN 与 5G 关键技术的结合能够进一步提升网络性能。参考文献[23]研究了大规模 MIMO 对 C-RAN 容量的提升，并通过构建部分集中的 BBU 池减少前传链路开销。参考文献[24]提出在 RRU 配备射频感知功能下，C-RAN 能够实现基于小区站址的射频资源分配，完成干扰消除提升网络容量。参考文献[25]研究了 C-RAN 与边缘缓存的结合，并比较分析了缓存配置对 C-RAN 性能的影响，包括前传链路负载与能量开销。

（2）雾无线接入网络（F-RAN）

相较于云计算，雾计算更靠近终端与用户。雾计算泛指网络边缘设备所提供的分布式计算能力，最早由 Cisco 公司[26]和普林斯顿大学著名学者 Mung Chiang 提出。为解决传统无线接入网络性能瓶颈，参考文献[27]将“雾计算”引入无线接入网络，提出了借助用户和边缘设备的计算及缓存潜能，缓解前传/回传链路压力的雾无线接入网络（fog RAN，F-RAN），C-RAN 和 F-RAN 对比



(a) C-RAN　　(b) F-RAN

图 2　5G 接入网络架构：C-RAN 与 F-RAN

如图 2 所示。在该网络架构中，网络边缘分散的计算和缓存能力能够通过协作进行有效整合，增强了本地实时处理、传输和管控能力，并通过承载下沉网络功能和边缘应用，实现本地信息处理和业务分发，提供更小的端到端时延性能。

F-RAN 由于其网络架构特征能很好地满足 5G 多样化的性能要求，一经提出就引起了业界的广泛关注，业界对 F-RAN 的各个方面开展了大量的跟踪研究。考虑到 5G 对时延的特殊需求，参考文献[28]在前传/回传链路容量受限情况下，分析了F-RAN中边缘缓存和前传容量对传输时延的影响，借助信息论揭示了网络性能与资源开销间存在的权衡关系，并设计了一种联合编码缓存和信号压缩方法。参考文献[29]对 F-RAN 移动性管理和资源管理进行了研究，设计与提出 F-RAN 自适应接入节点选择机制和干扰消除方法，能够减少异常切换发生概率和信令开销，提高网络资源利用率，验证了 F-RAN 作为 5G 无线接入网络架构的优越性。参考文献[30]调研了边缘缓存对网络性能，包括谱效、能效、时延的影响，提出一种 F-RAN 的自适应接入模式选择机制，在提供低传输时延的前提下，提高网络频谱和能量效率。

除了 F-RAN 本身的架构优势之外，F-RAN 能够很好地和 5G 关键技术相兼容，如大规模 MIMO、正交多址技术可以直接应用于 F-RAN 中。F-RAN 通过 SDN 实现了接入网络层面的 NFV。参考文献[31]提出了基于 NFV/SDN 的 F-RAN 架构，缓解了集中 SDN 控制器和骨干传输网的压力，实现了高效的编排管理。参考文献[32]提出了基于网络切片的 F-RAN 架构，将接入网络切片成多个逻辑独立网络，以满足不同业务差异化的性能需求。参考文献[33]提出了一种面向 5G 增强现实（AR）的 F-RAN 架构，将 AR 业务分解成多个可并行执行的子任务，并交由网络边缘节点进行分布式计算。参考文献[34]提出了一种面向差异化时延要求的 F-RAN 架构，借助通用处理器设备组建的仿真平台，实现并验证了对不同优先级业务的分级响应和处理，能够适配不同时延和速率要求的业务。

（3）接入网络切片（RAN slicing）

网络切片通过网络虚拟化技术，将网络中的各类物理资源抽象成虚拟资源，并基于指定的网络功能和特定的接入网技术，按需构建端到端的逻辑网络，提供一种或多种网络服务。由于核心网虚拟化程度较高，现有网络切片大部分都针对核心网。但是，探索研究接入网络中进行切片依旧存在必要性：考虑到 5G 的低时延、高容量与大连接的需求，接入网络中进行切片能够更好地提供差异化性能；网络切片作为端到端的逻辑网络，接入网络切片有助于补充完整现有的基于核心网

络的切片方法。参考文献[35]总结接入网络切片存在的挑战，并提出一种基于 LTE 的接入网络切片架构，在该架构中，接入网络被模块化，物理资源被抽象化，以实现灵活的服务编排与生命周期管理。参考文献[36]提出一种基于 3GPP eDECOR 的接入网络切片架构，并进一步地以 eMBB、uRRLC 和 mMTC 为例，借助 Open Air Interface 平台验证所提架构的有效性。

考虑到 5G 新特性，参考文献[37]研究 5G RAN 中，服务可解析成不同层空口协议栈的描述信息，并根据各层描述信息选择合适的协议栈配置，完成接入网络切片实例化。在 C-RAN 架构中，参考文献[38]研究了面向多租户环境的网络切片方法，并针对网络中多维度资源联合分配问题，提出分层的求解算法，保证用户传输速率需求的前提下最大化租户总收益。F-RAN 架构下，参考文献[39]探索了网络切片和边缘计算的融合，提出了基于边缘计算的接入网络切片架构，并指出资源管理和信息感知在该架构中的关键作用。

4 挑战和待研究内容

为了实现 5G 的“增强宽带，万物物联”，业界提出许多关键技术，5G 技术在提升网络性能的同时，也给 5G 系统架构带来了额外的挑战。例如，5G 性能需求的维度增加，5G 空口设计时需要在不同的性能指标之间进行权衡折中与联合优化，而为了应对多种业务共存而引入的网络切片技术将系统设计的复杂性进一步增加。可以预见的是，随着 5G 技术应用范围的扩展，5G 系统的复杂度也将呈指数型增长，系统设计和优化需要联合考虑不同域的技术的协同融合。

在传统方法无法对 5G 系统中存在的问题进行建模求解的场景下，人工智能技术的引入能够提供有效的帮助。例如，根据业务种类与网络环境，在为用户分配无线资源块时，网络资源调度器灵活选择合适的带宽与符号长度；基于用户和终端的上下文与位置信息，网络提供差异化的移动性管理；利用强化学习方法优化网络功能部署位置，完成网络的智能优化与管理。2017 年 11 月，ITU 成立了“机器学习”焦点组，重点研究机器学习、人工智能在包含 5G 系统的未来网络中的应用。随后，ETSI 发布了《自动化下一代网络中的网络和服务操作的必要性和益处》白皮书，强调 5G 网络中服务管理、运营自动化的目标。学术界也对 5G 和人工智能的结合展示了极大的研究热情。参考文献[40]首先总结了现有 5G 技术在网络资源、管理移动性管理等方面所透露的初步人工智能特性，探讨了人工智能在 5G 中进一步发展的必要性。

尽管人工智能在 5G 网络中的应用已经得到初步的探讨，神经网络与深度学习方法更是得到了仿真的初步验证，但 5G 网络与人工智能的结合依然处于初步阶段，未来仍旧需要进一步深化该方面的研究，包括但不限于人工智能算法中数据的获取与模型的选择以及 5G 网络在传输、存储和处理大数据时，网络成本和性能的折中。

5 结束语

本文对 5G 网络架构的研究与标准化进行了调研。在阐述 5G 愿景与需求、总结 5G 接入网络架构特征的基础上，从学术界和产业界分别介绍了 5G 接入网络架构的设计原理和具体组成，并根据现有工作，探讨了接入网络架构当前存在的挑战和对网络发展前景的展望，指出未来网络智能化的需求。

参考文献：

[1] 3GPP. Study on new radio access technology: radio access architecture and interfaces: TR38.801[S]. 2016.

[2] 董爱先, 王学军. 第 5 代移动通信技术及发展趋势[J]. 通信技术, 2014, 47(3): 235-240.

DONG A X, WANG X J. Technologies and future development trend of 5G mobile communication system[J]. Communications Technology, 2014, 47(3): 235-240.

[3] ITU-R. IMT-vision-framework and overall objectives of the future development of IMT for 2020 and beyond[R]. 2015.

[4] ANDREWS J G. What will 5G be?[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(6): 1065-1082.

[5] ETSI. Network functions virtualisation (NFV): architecture framework, document GS NFV002 v1.1.1[R]. 2013.

[6] 3GPP. Study on architecture for next generation system: TR23.799[S]. 2016.

[7] YOUSAF F, BREDEL M, SCHALLER S, et al. NFV and SDN—key technology enablers for 5G networks[J]. IEEE Journal on Selected Areas in Communications, 2018, PP(99): 1.

[8] TZANAKAKI A. Wireless-optical network convergence: enabling the 5G architecture to support operational and end-user services[J]. IEEE Communication Magazine, 2017, 55(10): 184-192.

[9] NGMN Alliance. 5G white paper[R]. 2015.

[10] PATEROMICHELAKIS E. Service-tailored user-plane design framework and architecture considerations in 5G radio access networks[J]. IEEE Access, 2017(5): 17089-17105.

[11] TALEB T, KSENTINI A, JÄNTTI R. Anything as a service for 5G mobile systems[J]. IEEE Network, 2016, 30(6): 84-91.

[12] VILALTA R. TelcoFog: a unified flexible fog and cloud computing architecture for 5G networks[J]. IEEE Communication Magazine, 2017, 55(8): 36-43.

[13] MARKAKIS E. Computing, caching, and communication at the edge: the cornerstone for building a versatile 5G ecosystem[J]. IEEE Communication Magazine, 2017, 55(11): 152-157.

[14] 3GPP. Proposal for study on a next generation system architecture: S2-153703[S]. 2015.

[15] 3GPP. System architecture for the 5G system: TS23.501[S]. 2018.

[16] 5GPPP. View on 5G architecture[R]. 2017.

[17] 3GPP. Way forward on the overall 5G NR work plan[S]. 2017.

[18] 3GPP. NG-RAN; architecture description: TS38.401[S]. 2018.

[19] LIN Y, SHAO L, ZHU Z, et al. Wireless network cloud: architecture and system requirements[J]. IBM Journal of Research & Development, 2010, 54(1): 1-12.

[20] SUNDARESAN K. FluidNet: a flexible cloud-based radio access network for small cells[J]. IEEE/ACM Transactions on Networking, 2016, 2(2): 915-928.

[21] TANG J, WEN R, QUEK T, et al. Fully exploiting cloud computing to achieve a green and flexible C-RAN[J]. IEEE Communication Magazine, 2017, 55(11): 40-46.

[22] BEYENE Y D, JANTTI R, RUTTIK K. Cloud-RAN architecture for indoor DAS[J]. IEEE Access, 2014(2): 1205-1212.

[23] PARK S, CHAE C, BAHK S. Large-scale antenna operation in heterogeneous cloud radio access networks: a partial centralization approach[J]. IEEE Wireless Communications, 2015, 22(3): 32-40.

[24] MEERJAK, SHAMI A, REFAEY A. Hailing cloud empowered radio access networks[J]. IEEE Wireless Communications, 2015, 22(1): 122-129.

[25] CHEN M, SAAD W, YIN C, et al. Echo state networks for proactive caching in cloud-based radio access networks with mobile users[J]. IEEE Transactions on Wireless Communications, 2017, 16(6): 3520-3535.

[26] BONOMI F, MILITO R, ZHU J, et al. Fog computing and its role in the internet of things[C]//Workshop on Mobile Cloud Computing, MCC'12, August 17, 2012, Helsinki, Finland. New York: ACM Press, 2012: 13-16.

[27] PENG M, YAN S, ZHANG K, et al. Fog-computing-based radio access networks: issues and challenges[J]. IEEE Network, 2016, 30(4): 46-53.

[28] TANDON R, SIMEONE O. Harnessing cloud and edge synergies: toward an information theory of fog radio access networks[J]. IEEE Communication Magazine, 2016, 54(8): 44-50.

[29] ZHANG H, QIU Y, CHU X, et al. Fog radio access networks: mobility management, interference mitigation, and resource optimization[J]. IEEE Communication Magazine, 2017, 24(6): 120-127.

[30] PENG M, ZHANG K. Recent advances in fog radio access networks: performance analysis and radio resource allocation[J]. IEEE Access, 2016(4): 5003-5009.

[31] LIANG K, ZHAO L, CHU X, et al. An integrated architecture for software defined and virtualized radio access networks with fog computing[J]. IEEE Network, 2017, 3(1): 80-87.

[32] XIANG H, ZHOU W, DANESHMAND M, et al. Network slicing in fog radio access networks: issues and challenges[J]. IEEE Communication Magazine, 2017, 55(12): 110-116.

[33] SHIH Y, CHUNG W, PANG A, et al. Enabling low-latency applications in fog-radio access networks[J]. IEEE Network, 2017, 31(1): 52-58.

[34] KU Y, LIN D, LEE C, et al. 5G radio access network design with the fog paradigm: confluence of communications and computing[J]. IEEE Communication Magazine, 2017, 55(4): 46-52.

[35] CHANG C, NIKAEIN N. RAN runtime slicing system for flexible and dynamic service execution environment[J]. IEEE Access, 2018(6): 34018-34042.

[36] KSENTINI A, NIKAEIN N. Toward enforcing network slicing on RAN: flexibility and resources abstraction[J]. IEEE Communication Magazine, 2017, 5(6): 102-108.

[37] FERRUS R. On 5G radio access network slicing: radio interface

protocol features and configuration[J]. IEEE Communication Magazine, 2018, 56(5): 184-192.

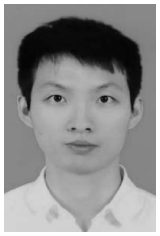
[38] HA V, LE L. End-to-end network slicing in virtualized OFDMA-based cloud radio access networks[J]. IEEE Access, 2017(5): 18675-18691.

[39] 项弘禹, 肖扬文, 张贤, 等. 5G 边缘计算和网络切片技术[J]. 电信科学, 2017, 33(6): 54-63.
XIANG H Y, XIAO Y W, ZHANG X, et al. Edge computing and network slicing technology in 5G[J]. Telecommunications Science, 2017, 33(6): 54-63.

[40] LI R. Intelligent 5G: when cellular networks meet artificial intelligence[J]. IEEE Wireless Communications, 2017, 24(5): 175-183.

[作者简介]

项弘禹（1993−），男，北京邮电大学泛网无线通信教育部重点实验室博士生，主要研究方向为雾无线接入网中切片架构及理论性能。



张欣然（1992−），男，北京邮电大学泛网无线通信教育部重点实验室博士生，主要研究方向为车联网和 5G 系统级仿真。

朴竹颖（1994−），女，北京邮电大学泛网无线通信教育部重点实验室硕士生，主要研究方向为无线接入网中的边缘缓存管理和优化。

彭木根（1978−），男，北京邮电大学网络技术研究院副院长、教授、博士生导师，主要研究方向为雾无线接入网络、后 5G 组网理论和关键技术、空间信息网络等。

专题：5G

面向通信与计算融合的 5G 移动增强/虚拟现实

周一青[1,2,3]，孙布勒[1,2,3]，齐彦丽[1,2,3]，彭燕[1,2,3]，刘玲[1,2,3]，
张志龙[4]，刘奕彤[4]，刘丹谱[4]，李兆歆[1,2,3]，田霖[1,2,3]
（1. 中国科学院大学，北京 100049；2. 中国科学院计算技术研究所，北京 100080；
3. 移动计算与新型终端北京市重点实验室，北京 100190；4. 北京邮电大学，北京 100876）

摘　要：面向通信与计算融合的 5G 移动通信网络，剖析移动 AR/VR 信息处理和传输的特征，提出融合通信与计算的能力，在未来 5G 移动通信网络中采用多级计算，通过协同网络多级计算节点的能力来解决移动终端计算能力有限的问题；采用智能传输机制，通过高效频谱感知、分层编码、空口自适应传输等来克服移动信道传输能力不稳定的问题；采用时延保障机制，通过通信与计算资源协同管理来确保移动 AR/VR 服务时延。对移动 AR/VR 多级计算模型、智能传输机制和服务时延保障的研究现状和发展趋势进行综述，总结现有研究存在的问题，并提出下一步研究的方向。

关键词：通信与计算融合；5G；增强/虚拟现实；多级计算
中图分类号：TN911
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018241

Mobile AR/VR in 5G based on convergence of communication and computing

ZHOU Yiqing[1,2,3], SUN Bule[1,2,3], QI Yanli[1,2,3], PENG Yan[1,2,3], LIU Ling[1,2,3],
ZHANG Zhilong[4], LIU Yitong[4], LIU Danpu[4], LI Zhaoxin[1,2,3], TIAN Lin[1,2,3]
1. University of Chinese Academy of Sciences, Beijing 100049, China
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
3. Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing 100190, China
4. Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: Oriented to 5G mobile communication network combining with communication and computation, the characteristics of mobile AR/VR information processing and transmission were analyzed, and the ability to integrate communication and computation was proposed, and the capability of integrating communication and computing was proposed. In the future 5G mobile communication network, multi-level computing will be used to solve the problem of limited computing capacity of mobile terminals by the capability of multi-level computing nodes in collaborative network. Intelligent transmission mechanism could overcome the problem of unstable transmission capacity of mobile channel by efficient spectrum sensing, layered coding and adaptive space port transmission, and the mobile AR/VR service delay was ensured by using delay guarantee mechanism and cooperative management of communication and computing resources. The research status and development trend of mobile AR/VR multi-level computing model, intelligent transmission mechanism and service delay guarantee was reviewed, the existing problems were summarized and the next research direction was put forward.

Key words: convergence of communication and computing, 5G, AR/VR, hierarchical computing

收稿日期：2018−07−01；修回日期：2018−08−10

1 引言

移动通信的发展日新月异，目前 5G 的研发已拉开大幕。相比 4G，5G 将在数据传输速率、传输时延、网络容量等多个方面实现飞跃性突破：峰值速率将达到 10 Gbit/s；支持超低时延超高可靠的服务，业务时延小于 5 ms；联网移动设备数量增加到现在的 100 倍，网络容量将提升 1 000 倍[1]。这些性能的突破增强了 5G 的智能业务服务能力。高通、ABI 等公司指出移动增强现实/虚拟现实（augmented reality/virtual reality，AR/VR）将成为 5G 的第一波杀手级应用[2]。

AR/VR 以计算技术为核心，生成逼真的视觉、听觉等，构成一定范围内的虚拟环境，用户可以与虚拟环境中的物体交互，获得身临其境的感受和体验。以 360°全景 VR 视频为例，其处理和传输流程如图 1 所示。为了增强用户的体验，首先用摄像机拍摄超高清分辨率的画面，然后拼接成 360°全景画面，给用户选择观看视角的自由，让用户具有身临其境的视觉体验。生成的 AR/VR 可缓存于服务器端，当用户端发出请求时，可通过有线或者无线传输将相应的视频提供给用户。显然，由于超高清分辨率和全景画面需要的多角度信息，生成的 AR/VR 相比普通视频，信息量可能高出几十倍，从而对系统的处理和传输能力提出很高的需求。另一方面，在用户端，当朝向、视角等状态发生变化时，系统可基于传感器等，运用动作捕捉技术，追踪用户行为并完成与虚拟场景的实时交互，实现更极致的互动体验。即用户发出新的画面需求，系统进行实时响应，在 20 ms 内将相应的画面提供给用户。由于用户的状态是随机改变的，所对应的画面通常是难以预存的，需要根据已有画面和变化的用户请求，渲染绘制出新内容，再提供给用户。渲染是一个复杂的综合性任务，通常需要借助深度信息计算、图像语义理解（也称作图像语义分割）等计算密集型处理来合成以假乱真的内容。可见，AR/VR 具有海量信息、密集计算的特点，它们对移动网络提出了大带宽、低时延的服务需求，为 5G 及未来移动网络的发展带来新的挑战和机遇。

AR/VR 在发展的早期多用于军事、航空航天、工业仿真等行业领域。近几年来，随着 Oculus Rift 等消费级 VR 产品的推出，AR/VR 逐渐向个人应用领域蓬勃发展。目前的 AR/VR 主要基于个人电脑（personal computer，PC），交互设备如头戴显示设备（以下简称头显）采用有线的方式与 PC 直连，计算处理任务由 PC 完成并通过连接线传回头显。受限于有线连接，用户不能自由活动，极大地影响了 AR/VR 的体验效果。因此，通过无线方式连接的移动 AR/VR 成为近年的发展焦点。2017 年 HTC

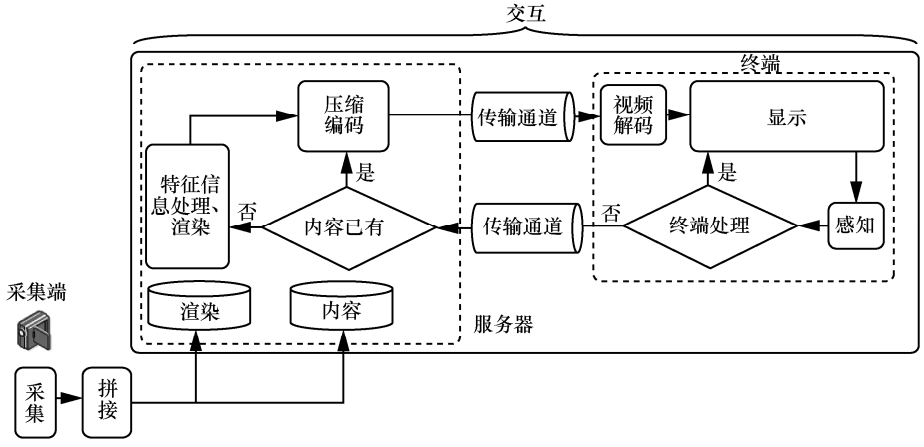


图 1 360°全景 VR 视频的处理和传输流程

发布了 VIVE 无线一体机，将头盔与 PC 之间的多根数据线升级为无线方式连接，但仍然需要配套一台高性能的 PC，成本高昂。而以谷歌 Daydream 和三星 Gear VR 为代表的联合手机终端的头显则用手机终端提供显示、通信和简单的计算功能，头显提供封闭的环境、镜头和交互功能。通过手机终端进行联网互动的同时将复杂的 AR/VR 处理任务传输到云端处理并将结果传回手机终端显示给用户。这种服务模式不需要配套性能强劲的 PC 来实现复杂任务的处理，能以低廉的价格提供移动 AR/VR 服务，正在逐渐成为移动 AR/VR 业务的主流实现方式。根据德意志银行的报告，未来几年移动 AR/VR 产品将赶超基于 PC 的 AR/VR 产品[3]。

2　移动 AR/VR 面临的挑战

目前基于手机终端的移动 AR/VR 仅能提供简单有限的体验，整体效果差强人意。如前所述，移动 AR/VR 信息的处理过程中，涉及渲染等高复杂度的密集型计算任务，而手机终端自身的计算能力有限，因此在终端处理任务的比例非常低，大部分任务都需要通过移动互联网传送到云端服务器计算并传回给用户，时延大。另一方面，在移动 AR/VR 信息的传输过程中，受限于移动通信的传输带宽以及信道质量不稳定等问题，导致用户接收到的视频画面质量不稳定、不流畅，体验效果差。因此，移动 AR/VR 的信息处理和传输面临众多挑战，需要解决图 2 所示的三大矛盾。

（1）移动 AR/VR 处理的密集计算需求与有限的移动终端计算能力之间的矛盾

移动 AR/VR 具有的海量信息导致处理复杂度高，对系统的密集计算能力提出了很强的需求。以典型的多视角三维场景重建为例，在较低分辨率的摄像头下支持 320 像素×240 像素，在约 1 m×1 m×1 m 真实环境空间中使用 256 体素×256 体素×256 体素的体素解析度进行计算，通过图形处理器（graphics processing unit，GPU）并行处理也只能达到 15 帧/s 左右的交互式帧率（电脑配置：计算机处理器 i7，32 GB RAM，NVIDIA Titian black 显卡，6 GB 显存）；若提升摄像头分辨率，且在更大的真实环境中进行应用，那么需要采用更高密度的体素辅助计算，如果体素解析度提高至 1 024 体素×1 024 体素×1 024 体素，计算量又将提高 64 倍左右。相比 PC 配置，移动终端虽然逐步配备了多核处理器和图形处理核等，其计算能力还存在明显差距，难以满足移动 AR/VR 密集计算的需求。若将计算迁移至远端云服务器进行，则其时延性能难以保障。值得注意的是，虽然移动终端本身计算能力有限，如图 2 所示，已有研究[4]指出，未来移动网络将在基站、核心网等网络的不同层次融合移动边缘计算（mobile edge computing，MEC），成为一个通信与

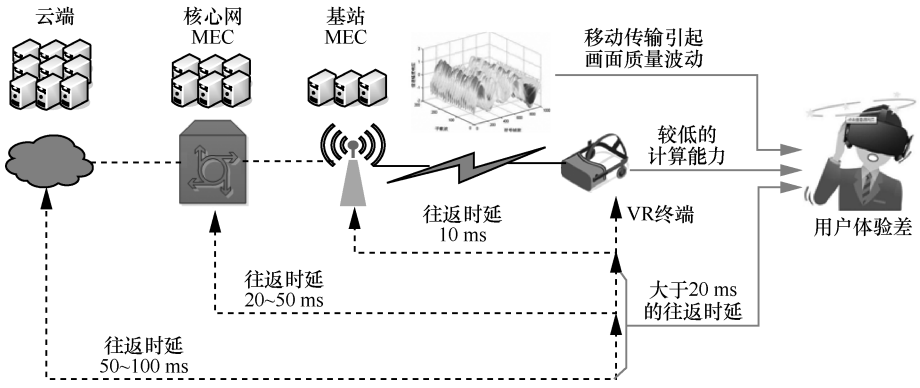


图 2　移动 AR/VR 处理与传输问题分析

多级计算协同融合的网络。若能有效协同移动终端与未来 5G 移动网络的多级计算能力，有可能完成 AR/VR 的密集计算需求。

（2）移动 AR/VR 的高画质需求与随机时变的移动传输能力带来的画质不稳定之间的矛盾

为了增强用户的体验，需要提升移动 AR/VR 的清晰性、流畅性，因此对传输能力也提出了大带宽、高质量的需求。相比有线通信，采用无线传输的移动通信可以在任何时间、任何地点为任何人提供服务，这种服务的自由性提升了用户的体验，但无线传输信道不光带宽有限，而且其信道质量随着时间和地点不断动态变化，这种随机的时变传输能力为 AR/VR 传输带来了很大的问题。一方面，AR/VR 视频流具有大象流特征，对带宽需求特别高；另一方面，鉴于 AR/VR 的交互特性，用户对视频的需求存在强烈的随机性和突发性，即便是看同一个 AR/VR 视频，不同的用户视角不同，也会发出不同的画面请求，进一步增大传输带宽的需求。当用户发起新的画面需求时，如果当时信道质量较差，只能传输低质量画面，引发的视觉疲劳会带来眩晕感，如果以低速传输高质量画面，可能无法及时响应用户需求，导致时延过大或者画面停顿，同样严重影响用户体验。这个问题在通信与多级计算融合的未来无线网络中有可能得到解决。未来无线网络引入的 MEC 服务器，具有大数据计算和存储的能力。一方面可以通过无线大数据挖掘，获得频谱地图，当移动信道质量出现下降时，快速提供可切换或可扩展的频谱，提升传输能力。另一方面，正如已有研究[5]指出的，协同利用通信与计算资源，有可能设计智能高效的移动传输机制，完成传统移动通信网络中难以传输的移动 AR/VR。因此，解决第二个矛盾需要针对移动 AR/VR 和随机时变的移动无线信道的特征，研究通信与计算协同的智能传输机制，确保移动 AR/VR 传输画面的质量和流畅性。

（3）移动 AR/VR 的低时延需求与现有移动网络服务时延过大之间的矛盾

AR/VR 业务一般需要低至 20 ms 的服务时延来保障其自然、流畅的体验，而极致的体验，需要进一步降低服务时延。服务时延包括业务的传输时延和处理时延。如图 2 所示，传输时延部分，现有 4G 网络的移动终端与基站空口传输往返时延（round trip time，RTT）近似为 10 ms；如果需要云端的计算能力，由于终端到云端的数据传输需要经过网关等设备经过多重转发，往返时延一般为 50~100 ms，远大于移动 AR/VR 要求的 20 ms 的服务时延要求，严重影响用户的体验。另一方面，移动 AR/VR 的处理时延由计算节点的计算能力及业务的计算量决定。目前移动终端计算能力较弱，而云服务器计算能力则非常强大，故相同计算量的业务在移动终端和云服务器的处理时延差异较大；同时，移动 AR/VR 的海量数据也带来了非常大的计算量，以 4K RGB（red-green-blue）360°全景视频为例，其全视角分辨率需达到 7 680×3 840 像素，是普通 720P 视频的 32 倍，处理和传输时延都将大幅增加。可见，目前移动网络以移动终端协同远端云服务的方式提供 AR/VR 业务，很多时候仅传输时延就超出了 AR/VR 的时延限制。但如果考虑未来通信与多级计算融合的移动网络，一方面 5G 的空口时延将降低到 1 ms 左右，另一方面有可能通过协同 MEC 边缘计算节点完成 AR/VR 的密集计算任务，将大幅降低传输时延。而 MEC 服务器相比云服务器，计算资源相对有限，其处理时延不可忽略。因此，针对移动 AR/VR 的需求，基于未来移动网络中的传输和处理时延模型，对通信及计算异质资源的分配进行优化，有可能降低总体服务时延，满足移动 AR/VR 服务的时延需求。

综上所述，为了解决当前移动 AR/VR 面临的三大矛盾，面向未来 5G 移动网络，以通信与计算融合为核心思想，应紧密结合移动 AR/VR 业务特

征与需求，研究未来 5G 移动网络的多级计算网络模型及移动 AR/VR 信息处理的多级协同计算方法，研究多级计算架构下移动 AR/VR 的智能传输、服务时延模型及时延优化与保障机制，提出基础理论和核心机制，解决移动 AR/VR 信息处理与传输的关键问题。下文将对移动 AR/VR 的“计算模型”“智能传输”和“时延保障”3 个方面的研究现状及发展动态进行综述。

3　AR/VR 信息处理计算模型与方法

AR/VR 信息处理过程中，需要实时完成场景深度估计、图像语义理解、三维场景重建、高真实感渲染等密集计算型任务，以保障用户可以获得自然、流畅的体验。目前 AR/VR 的信息处理研究主要是通过并行计算、分布式计算完成密集计算型任务。此外，针对移动终端有限的计算能力难以支持移动 AR/VR 密集计算任务的问题，研究人员也提出了终端—服务器的两级计算方法和面向移动终端的一系列低复杂度算法。

（1）AR/VR 的并行计算与分布式计算

在 AR/VR 应用中，考虑到 PC 的 CPU 计算资源有限，计算领域提出了 GPU 并行计算、多 PC 分布式计算的方法，通过增加计算资源的方式完成海量信息处理需求。在 GPU 并行计算方面，以 AR/VR 三维场景重建为例，Dame 等[6]提出了在具有两个 GPU 的 PC 上进行基于并行加速的三维场景重建，首先将三维场景重建拆分为深度估计与目标位姿估计、目标分割及关键帧提取等多个子任务，然后将它们分配到不同 GPU 上并行执行。Pradeep 等[7]提出了在具有一个 GPU 的 PC 上进行三维场景重建的并行处理方法，首先将场景重建按像素拆分为若干子任务，然后在 GPU 的多个线程上并行处理各子任务，实现快速计算。上述研究表明，AR/VR 业务的海量信息处理包含多个任务，如三维场景重建、高真实感渲染，部分任务之间可以进行并行处理，同时部分任务还可进一步分割为多个子任务，如三维场景重建可拆分为深度估计、关键帧提取等子任务实现并行加速；但上述任务或子任务并行处理方法均是基于单一 PC 上的 GPU 并行计算的，故其任务分割方案具有局限性。

区别于参考文献[6-7]只基于一个 PC 计算节点进行并行计算，Lan 等[8]提出了一种在分布式计算机系统中实现三维场景重建的并行处理机制，其主要思想是利用不同计算机并行完成视频帧的去噪滤波及深度传感器校验，然后通过吉比特局域网将视频帧传输给主服务器进行数据融合。Paweł 等[9]研究了在具有不同计算能力的计算节点上进行并行处理的机制：首先在具有 NVIDIA Tesla S1070 与 NVIDIA Tesla C2070 两种共 4 个 GPU 的 Hal 机上进行 3D 图像重构，利用电容层析成像技术，选择输入矩阵的不同行对应的某一电极并将其分配给特定的 GPU 实现并行计算；其次，增加具有 2x GTX 570 GPU 的 Dave 机，仍采用上述任务分配方法，研究不同计算节点的并行计算。上述在分布式计算机系统进行并行计算的模型具有可扩展性，即可以通过增加计算节点的方式来完成更大规模的计算任务。但模型中各节点计算能力虽然不同，却基本相当，且计算任务对时延没有严苛要求。而通信与计算融合的未来 5G 移动网络将包含手机移动终端、PC、服务器等计算能力存在数量级差异的节点，同时传统基站、集中式接入网（centralized radio access network，C-RAN）设备[10]、终端直通（device to device，D2D）[11-12]等共存的异构通信网络环境时，单个性能差的计算节点或传输环境都可能导致整个业务出现不可预知的延迟。

（2）移动 AR/VR 终端—服务器两级计算方法

在移动 AR/VR 方面，由于手机终端配备的计算能力逐步提升，如多核处理器和图形处理核，同时配备微电子传感器，能够处理角速度和线性加速度，为移动 AR/VR 信息处理提供了新的可

能。Alex 等[13]提出了一种移动终端与云服务器联合进行三维场景重建的方法，其中移动终端进行实时跟踪、关键帧选择及低复杂度关键帧更新，而具有深度信息的关键帧则合并到位姿图中传输给云服务器进行三维场景重建。该模型中云服务器作为移动终端协作网络的中心节点，能够同时支持大量移动终端的密集计算需求，同时用户的三维模型可以存储在云端，被其他用户复用或在此基础进行完善，有效节约计算资源。但参考文献[13]将计算任务卸载到云端进行计算，需要经过路由器、网关等层层关键设备，其交互时延难以得到保证。

针对移动终端计算能力难以满足移动 AR/VR 密集计算需求，而云端计算难以保证时延的矛盾，雾计算（fog computing）[14-16]、朵云（Cloudlet）计算[17]、MEC [18]等概念相继提出，旨在移动网络边缘提供 IT 服务环境和云计算能力，使业务处理更靠近终端，在满足密集计算要求的同时，有效降低业务延迟。Hou 等[19]提出将便携式 VR 眼镜与云/边缘计算设备通过无线进行连接，分别将服务器部署在云端、网络边缘（移动网络网关、基站或无线接入点）、终端设备边缘，利用服务器的计算能力来减轻 VR 眼镜重量，分析了不同部署策略适应的业务类型。但该研究主要关注利用视频编码、传输及压缩方法来降低传输视频流比特率，并未解决云/边缘计算服务器如何辅助终端设备进行密集计算的问题。

（3）面向移动终端的低复杂度 AR/VR 处理方法

Tanskanen 等[20]提出了一种在手机终端上进行稀疏三维重建的算法，利用手机终端的惯性传感器进行实时跟踪与建图，通过选择合适的关键帧、估计场景重建尺度，实现了一组基于稀疏点云的高效三维重建。Kolev 等[21]提出了一种在手机端进行基于深度图的三维场景重建方法，通过评估局部几何方向、底层相机设置和光度等来确定每个深度估计的权重，进而实现深度信息融合。由于手机终端计算能力有限，只能实现稀疏三维场景重建，与稠密重建质量仍有较大差距，影响用户体验。此外，为了在移动终端获得混合现实的体验，需要算法能够实时顽健地感知场景的三维信息。Ondruska 等[22]提出一种可在主流手机上运行的实时场景估计方法，但是由于需要基于体素的深度融合，限制了该方法只能在较小的场景下使用。Mur-Artal 等[23]提出基于 ORB（oriented FAST and rotated BRIEF）特征的相机定位和稀疏场景重建方法，该方法可移植到移动端运行，但是由于场景重建仅包含一些稀疏的三维点，无法满足混合现实的需要。此外，利用 AR/VR 视频视野范围很大的特点，Lai 等[24]提出将全景视频划分为规则的区块以实现在移动端并行的低复杂度视频解码，能显著减少时延，然而这种简单的图像分块没有考虑图像本身的内容，也没有考虑时序上的冗余，例如一些图像的内容要比另外一些具有更高的视觉显著度，因此处理它们的优先级应是不同的。

总结上述研究现状可知，现有 AR/VR 信息处理研究利用并行计算和分布式计算来应对其海量信息的密集计算需求，说明 AR/VR 的密集计算任务是可分割的，但这些计算模型和方法都是针对具有相同或相当计算能力的节点设计的，在通信与计算融合的未来 5G 移动网络中，计算节点能力与其部署的通信网络位置有关，可能存在数量级的差异，现有计算模型和方法无法直接应用，必须面向通信与计算融合的未来 5G 移动网络研究新的计算模型，并提出相应的协同计算方法。此外，目前虽然面向移动终端提出了一些低复杂度的 AR/VR 视频信息处理算法，但这些算法尚未充分利用 AR/VR 视频的时空特性，以牺牲精度为代价降低计算量，会给用户带来很大的不适。因此亟待解决的难点问题是如何在移动终端实现顽健轻量级移动 AR/VR 处理算法。

4 移动 AR/VR 的智能传输

为了解决移动网络时变的传输能力带来的移

动 AR/VR 服务质量不稳定的问题，主要从两个方面进行探讨：一是从系统的角度适配业务传输的需求，比如改变传输频带、增加传输带宽；二是从业务处理和传输的角度适应移动信道的变化，比如面向移动传输的视频分层编码、空口的自适应传输等。

（1）无线频谱感知与动态适配

提供更多更好的可用频谱是保障移动 AR/VR 传输需求的一个有效技术。从统计上看，移动 AR/VR 业务在不同的时刻数据传输量可能会发生剧烈的抖动，带宽需求具有快速时变特点。为了获得频谱，首先要进行频谱感知，主要包括节点频谱感知和频谱地图两种方式。节点频谱感知方面，已有研究针对未来移动网络，提出了基于认知导频信道的快速频谱感知方法[25]，但此类方法普遍存在实现简单但正确性低或者感知性能好但时延长的问题[26-27]。而频谱地图基于大量的计算和存储，能快速提供较为准确的频谱信息。参考文献[28]采用频谱地图，记录用户使用频谱的频段、时间、用户位置以及服务质量等多维信息，为精确、安全、快速的频谱感知提供参考。参考文献[29]论述了频谱地图的应用，将事件学习算法和知识学习算法融入频谱地图，在更好地实现各项功能的同时，有效保护授权用户。参考文献[30]中介绍了一种多级无线电环境地图的架构，在核心网侧、基站侧、终端侧均放置了计算处理单元，分别用于生成各层级的无线电环境地图，实现整个覆盖区域频谱资源的高效协调。移动 AR/VR 业务作为时延敏感的高带宽传输业务，对频谱质量、时延等性能提出全面的要求。但上述研究重点关注可用频谱信息，未考虑频谱质量问题，不能满足 AR/VR 获取高质量传输频带的需求。在频谱适配方面，现有研究主要考虑用户公平性、系统吞吐量等方面性能。参考文献[31]提出了跨成员载波的比例公平调度算法，可以保障载波聚合下多用户调度的公平性。参考文献[32-33]提出了基于贪婪算法的联合载波选择与资源块分配的资源调度方法以提高系统吞吐量。参考文献[34]提出了多载波共用一个缓存队列的联合队列调度方式，相比于独立队列调度方式可以充分利用分集增益。参考文献[35]分析了用户调度服从泊松分布时，基于轮询与移动散列算法的频谱适配的性能。值得注意的是，上述研究都未能与 AR/VR 结合。在 AR/VR 场景中，一个用户请求在处理时通常被分割成多个业务，如三维场景重建、高真实感渲染等，这些业务的数据传输量有很大差异。因此，在 AR/VR 应用时，需要考虑这些不同业务的用频需求差异，此外，还需要考虑同一业务不同编码层对服务质量的需求差异。因此，现有频谱适配无法满足 AR/VR 多种差异化业务的需求。

（2）面向移动传输的视频分层编码

在传统视频信息处理研究中，目前已经提出了使用不同的编码方式适应移动传输环境，为用户提供不同等级质量的视频服务。两种主流的编解码方案为可分级视频编码（scalable video coding，SVC）和感知视频编码。SVC 方面，主要是基于 H.264/AVC（advanced video coding）混合编码框架，在时间、空间、质量上实现分层编码，以实现不同帧率、图像分辨率和图像质量等级的自适应调整[36]。H.265/HEVC（high efficiency video coding）与 SHVC（scalable extension of HEVC）作为 H.264/AVC 的继任者，同样继承了分层编码的特性[37-38]。SHVC 支持不同空间分辨率或重建信噪比（signal noise ratio，SNR）的多层视频序列的编码，支持高达 8 个分层，包括一个基本层和多个增强层。通过引入上采样滤波、层间纹理预测、层间运动预测等技术，SHVC 还支持混合编解码可分级、比特深度可分级和色彩域可分级。感知视频编码方面，主要是利用了视频信号中存在的视觉冗余，即人眼不能察觉的图像中的某些信息。人类视觉系统的研究表明，人眼对图像和视频的感知是有选择性的，不同的对象或者区域

具有不同的视觉重要性，并且对视觉信号的各种失真具有不同的敏感和容忍程度。基于视觉特性的视频编码主要思路是如何根据视觉感知特性选择优化的编码参数，实现码率的优化分配，例如为视觉重要区域分配更多的码率资源来提高其主观质量；减少非视觉重要区域的码率资源来减少视觉冗余，从而提高编码性能[39]。此外，多视点视频能提供立体感和交互性，结合多视点视觉感知模型的特点，基于视觉特性的多视点视频编码可分为基于立体视觉注意的多视点视频编码和基于立体视觉可见度的多视点视频编码[40-41]。基于下一代编码标准 H.265/HEVC 的视频感知编码可能成为未来 AR/VR 编解码的重要工具[42-43]。由上述研究可见，当前视频编码大多面向传统视频，尚未有针对 AR/VR 的编解码机制。在 AR/VR 编解码研究中，应结合其独有的特征如深度信息、六自由度和用户的感兴趣区域的差异性等，一方面提高编码压缩比，轻量化业务数据量；另一方面需要降低编码复杂度，降低对移动设备的计算压力和续航压力，为智能地在时变的移动无线环境中提供不同的体验提供编码方案的基础。

（3）空口感知的 AR/VR 自适应传输

由于无线信道的时变特性，很难满足 AR/VR 持续大带宽和低时延的需求，移动 AR/VR 传输面临很大挑战。此外，考虑多用户场景，由于用户状态不同，即使观看同一个移动 AR/VR，所需的画面也不完全相同，无法像普通视频那样，采用多播的传输方式来提升频谱效率。这进一步加大了移动 AR/VR 的带宽需求。如何设计能够感知空口的变化、符合业务和无线网络特点的自适应机制是亟待解决的问题。现有 AR/VR 传输的成果大多基于有线网络环境，只有少量的相关工作基于无线环境。例如，在参考文献[44]中，运用博弈论等方法，提出了一种异构蜂窝网络下适合 AR/VR 传输的资源管理策略，但时延仍然难以达到 AR/VR 业务要求。参考文献[45]提出了一种时延导向的基于无线局域网的 AR/VR 多用户接入策略，分析了传输时延的组成部分，制定出了一套 AR/VR 多用户接入方案。另外一方面，云服务的引入为移动视频传输提供了新的思路。首先，可利用集中式特点来进行更好的资源管理。已有研究面向集中式移动网络架构，提出负载感知的资源管理，可提高资源利用率达 70%[11]。其次，可利用集中式架构中心处理单元的计算资源进行视频转码，根据用户需求将原始视频转码成较低版本，在空口传输时节省传输内容，提高空口资源利用率。在参考文献[46]中，提出基于集中式云服务的转码机制，利用云计算，以用户的信道条件所需的视频质量为参数来计算转码版本，提升视频质量和频带利用率。参考文献[47-48]针对一对多实时转码的视频直播服务，提出了一种云计算资源分配的方案，在提供用户所需的视频质量的前提下结合地域性差异最小化计算资源的开销。参考文献[49]则是在参考文献[47-48]的基础上，针对流式直播服务，为视频接收者和发布者设计了一套云服务器联合选择策略，降低了系统开销和地域性差异的影响。根据直播视频发布者受欢迎程度的不同，参考文献[50]借助计算资源，通过建立多目标优化问题，最小化计算资源开销和最大化用户体验，在时延约束的条件下利用李雅普诺夫优化的方法求解并提出了一套计算和带宽资源联合分配方案，使网络性能和用户体验得到了权衡。参考文献[51]将云计算与内容中心移动网络结合，为多媒体服务提出了一种较为完整的云网络架构，包括计算资源的部署和网络的拓扑结构。但以上研究尚未与移动 AR/VR 特征相结合。可见，目前针对移动 AR/VR 业务的传输策略研究尚处于起步阶段，在通信与计算融合的移动网络架构下对 AR/VR 传输策略的研究国内外均属空白，尚未见任何成果公布，其中蕴含着大量研究机会。针对移动 AR/VR 传输的资源分配、版本选择、多播等技术，都有必要探索新的解决思

路和方法，从而更为有效地改善移动 AR/VR 业务的用户体验。

总结上述研究现状可知，作为一种崭新的移动多媒体服务，面向 AR/VR 的移动传输机制研究目前尚处于初级阶段，在频谱感知、信源分层编码、空口自适应传输等都少有针对性的研究，亟需根据移动 AR/VR 的需求，剖析其与移动传输相关的特征，面向通信与计算融合的未来 5G 移动网络，设计智能的传输机制。

5　移动 AR/VR 服务时延模型及优化保障

用户交互得到即时响应是移动 AR/VR 良好用户体验的主要来源。当服务时延大于图像刷新时间间隔时，用户将产生晕眩感。因此，服务时延是影响体验的重要因素。目前在这个方向上已有大量的研究，提出了多种时延优化保障机制，研究的主要思路是基于给定的服务时延模型，从移动网络的资源管理、移动切换、AR/VR 用户位姿预测、边缘缓存等方向展开服务时延的优化或保障。

（1）移动 AR/VR 的服务时延模型

目前移动 AR/VR 服务时延模型相对比较简单，仅考虑单服务器节点情况。Deng 等[52]研究了在小区基站部署云服务器进行密集计算任务的卸载，其中假设应用的任务已在云服务器虚拟机上进行备份，故其服务时延模型不考虑任务的传输时延，仅考虑任务的处理时延以及各任务之间彼此依赖且上一任务与当前任务在不同节点进行处理时上一任务结果的传输时延；Chen 等[53]研究了将密集计算任务卸载到移动边缘云计算的时延模型，其主要包括任务在 3G/4G 无线信道的上行传输时延及任务在移动边缘云服务器的处理时延，而未考虑任务处理结果的下行回传时延；Lai 等[24]研究了手机终端 VR 业务卸载到台式机进行处理的时延模型，其中手机终端与台式机通过基于 IEEE 802.11ac 协议的无线局域网（wireless local area network，WLAN）进行相连，服务时延则主要包括请求等待时延、视频流原始帧的传输时延及在台式机计算时延，其中视频流原始帧的 WLAN 中的传输时延为主要时延；Ahn 等[45]研究了 WLAN 中基于 IEEE 802.11 协议的多用户 VR 服务时延模型，其服务时延主要包括用户感知时延、终端决策时延、上行传输时延、PC 处理时延、下行传输时延及终端整合时延，该模型中考虑了多用户场景下信道接入时延问题，重点分析了多用户业务上行数据分组传输时延。上述模型均是基于单服务器节点的服务时延模型，其结构简单，很难直接拓展到包含 D2D 通信、传统小区基站通信与 C-RAN 集中式架构等并存的异构通信网络，且异构通信网络中包含终端、MEC、云服务器等多级不同计算能力的服务节点，存在多种计算方式，同样影响移动 AR/VR 服务时延建模。

（2）基于资源管理的移动 AR/VR 的服务时延优化

在给定服务时延模型基础上，针对移动 AR/VR 服务时延的优化主要集中在 3 个方面，即研究时延约束下用户终端功率损耗问题、时延与功耗的联合优化以及直接以最小化服务时延为目标。在时延约束下最小化用户终端功耗方面：Cao 等[54]提出了单用户场景中基于组合优化算法的最优自适应算法和基于贪婪算法的次优算法，将终端能耗分别降低 48%和 47%；Zhao 等[55]假设所有用户具有相同的信道质量和计算能力，其提出的最优算法和次优算法能分别节约 40%、30%的能耗；Vondra 等[56]探讨了通信与计算负载的平衡，在满足时延要求的前提下提出一种应用考虑算法，根据当前计算和通信负载来选择合适的接入基站。在时延与功耗的联合优化方面：Muñoz 等[57]提出一种联合分配无线通信和计算资源的通用架构以对用户终端能耗和时延做折中处理；对于多用户场景，Mao 等[58]基于 Lyapunov（李雅普诺夫）优化提出一种在线算法用于决策计算任务的分配，

在每个时隙，利用 Gauss-Seidel（高斯—塞德尔）方法来决定计算卸载的优化传输功率及带宽分配。在最小化服务时延方面：Liu 等[59]采用马尔可夫决策过程方法对基与任务缓存排队状态、本地处理单元执行状态以及传输单元状态的计算任务进行调度，通过分析每个任务的平均时延及移动终端的平均功率损耗，研究功率约束下的最小化时延问题，并通过一维搜索算法寻找优化调度策略；Mao 等[60]研究了加入能量收集设备的 MEC 系统中单用户计算卸载策略，综合考虑了卸载决策、CPU 周期频率以及计算卸载的传输功率，以最小化时延为目标，通过 Lyapunov 动态优化算法实现优化目标。上述时延优化问题均假设待处理的任务是可被任意分割的，但是实际上，移动 AR/VR 任务的分割比例是与业务本身息息相关的；同时上述问题中仅针对终端及 MEC 或云端两级计算节点，且都假设网络中通信、计算与存储资源是有限的，但是并没有探究如何进行资源部署以及资源部署对服务时延的影响。

（3）计算节点移动切换机制

在通信与计算融合的未来 5G 移动通信网络中，通信与计算节点切换是必不可少的核心技术，切换时延对移动 AR/VR 的体验有重大影响[61]。计算节点的切换主要分为按需切换和主动推送两大类。按需切换方面，参考文献[62]提出分布式 MEC 服务器总是跟随服务基站无线切换而变化，可以实现业务传输时延最小化。参考文献[63]考虑协作小区通信方式，根据终端时延需求等确定服务小区，通过无线链路和终端相连，协作组内的其他小区和服务小区通过回传链路连接。通过马尔可夫决策过程（Markov decision process，MDP）算法选择最优切换节点，在满足能耗需求的条件下，最小化业务总时延。参考文献[64]考虑云、分布式 MEC 和 D2D 协作多级计算架构，在终端与基站连接时间等限制下，进行计算节点切换决策，最大化系统接入能力。已有计算节点按需切换机制普遍存在切换时延长的弊端，切换时延至少是 60 ms[65]，甚至可达数秒[62]，远不能满足移动 AR/VR 提出的 RTT 为 20 ms 的要求。主动推送切换方面，其核心思想是利用移动终端具备定位能力，其移动具有规律性和可预测性的特点，主动将计算任务提前推送到将要切换的目的节点，改善切换时延性能。参考文献[66]基于终端、MEC 和云端三层架构，通过挖掘终端历史数据，采用多项式非线性回归的方法，预测终端位置，提前进行 MEC 切换，保证服务的连续性。参考文献[67]假设可以采用一定的方法预测出未来 T 时间窗口内，每一个 MEC 执行任务的传输成本及业务在 MEC 间切换的成本上限，采用 MDP 方法，确定服务 MEC 序列，最小化终端在 T 时间的平均成本。参考文献[68]在参考文献[63]的基础上，加入终端位置预测，进行最优切换选择机制的设计，进一步降低业务处理时延和能耗。现有的主动推送切换机制大都是基于位置预测进行推送，没有考虑切换时机和新入网终端等因素。而移动 AR/VR 的主动推送切换需要切换的时间、地点、内容 3 方面的预测与实际需求精确匹配，才能保证内容推送的实时性和精确性。因此，仅基于位置预测的传统主动推送机制性能无法保证移动 AR/VR 业务切换的性能需求。

（4）基于用户位姿预测的时延优化机制

AR/VR 业务良好体验的重要来源之一就是自由的用户位姿，主要分为用户在虚拟现实中的运动任务、用户在空间中的自由移动以及基于空间位置的行为交互。无论何种用户行为，都需要大量的计算资源开销。基于用户行为的位姿降低 AR/VR 服务时延，提升实时计算性能，是目前业界研究的热点。Fang 等[69]提出了一种基于视觉惯性的实时用户位姿跟踪算法，结果表明该方法能够实时、稳定地为移动虚拟现实提供一种平滑、稳健的六自由度运动跟踪。Kim 等[70]使用事件相机（event camera）实现了基于用户位姿实时重建

立体场景的方法，基于 3 个解耦的概率过滤器，分别追踪不同自由度的用户位姿，计算场景强度的梯度图和场景的深度图，并合成为关键帧，通过时域和空域的超分辨率重建从低比特率的事件流中恢复实时三维场景视频序列。Dobbins[71]提供了一个运动捕捉和虚拟现实的协同可视化系统，确保了六自由度 AR/VR 业务与用户交互的可能。位姿捕获系统捕获一个或多个用户的头部旋转信息，从而控制真实世界视频的平移、倾斜和缩放。当用户头部的位置改变时，其所视的虚拟视频内容需要实时相应改变。以上方法分别从视觉惯性、运动捕捉等方面进行了用户位姿的跟踪及相应视频内容的播放，并没有考虑视频内容的兴趣区域。对于移动 AR/VR 而言，用户对视频内容兴趣区域的变化同样可以概括为用户位姿的范畴。此外，当前没有针对用户位姿预测的研究，由于移动 AR/VR 业务计算量大，会产生较大的业务处理时延，如果能够对下一时刻用户位姿进行预测，并根据预测结果进行计算资源的提前分配实现任务的提前计算，可有效降低处理时延。

（5）低时延缓存策略

如前所述，未来移动网络将融合 MEC 节点，提供计算与存储功能，可以使移动 AR/VR 内容更加靠近用户。如果用户请求的视频内容被缓存在其可接受服务范围内的缓存点中，那么用户可以直接从这些缓存点获取视频内容，不需要再经过核心网或从远端视频服务器获取内容。缓存命中时，服务时延可以得到有效降低[72]。目前，研究人员针对不同无线网络场景下的高效缓存策略进行了大量研究，通过分析视频业务的特点、用户历史观看行为和具体的网络架构，使用户所请求的视频内容能以更高的效率被缓存[73-74]。随着研究的深入，针对具体业务特点的缓存机制研究逐渐受到关注。例如，为降低业务时延，参考文献[75]基于内容请求变化率和文件流行度设计了缓存策略，可满足严格的时延要求；参考文献[76]参考内容分发网架构提出了分布式的自适应流媒体服务，能提高视频服务效率降低时延，但是文中缓存策略针对基于内容分发网络的超文本传输协议（hypertext transfer protocol，HTTP）动态自适应流媒体（dynamic adaptive streaming over HTTP，DASH）业务设计，不能直接用于 AR/VR 业务。参考文献[77]提出了预取与缓存整合策略，降低了比特率抖动问题，有效提升了视频流命中概率。为适应视频流媒体业务的可按照多种速率进行分发的特点，参考文献[78]基于最近最少使用内存清理提出通过最高速率视频段缓存和转码的结合，有效提升了缓存命中率。此外，参考文献[79]通过采用分层视频编码的视频传输，提高了缓存命中率。参考文献[80]提出了一种基于云服务的多级视频缓存架构，在边缘缓存的基础上，还在核心网部署云服务器和云缓存，在降低网络传输时延的同时提升了用户体验质量（quality of experience，QoE）。在缓存策略设计方面，已有以提高缓存命中率为目标，设计了面向 DASH 的云接入网缓存策略[81]。根据上述调研，目前针对业务特点的缓存研究已经逐步深入，但是尚未有研究充分结合未来移动网络通信与计算融合的架构特点以及移动 AR/VR 业务的多视角特点进行缓存策略设计。

总结上述研究现状可知，鉴于服务时延对移动 AR/VR 的重要性，目前已经展开了大量相关研究，但仍存在一系列挑战。在服务时延模型方面，现有基于单服务器节点的服务时延模型过于简单，必须面向通信与计算融合的未来 5G 移动网络架构，研究匹配的移动 AR/VR 服务时延模型；在基于资源管理的时延优化与保障方面，未能与移动 AR/VR 特征紧密联系，也缺乏从整个系统层面的资源保障，必须基于移动 AR/VR 处理与传输的特征，分析通信与计算资源对服务时延的影响并提出相应的优化保障机制，例如可以研究网络切片机制，从系统层面确保移动 AR/VR 的资源，从

而保障时延性能；在低时延缓存和基于用户位姿预测的时延优化机制方面，也有很大的研究空间。

6 结束语

虽然移动 AR/VR 有望成为未来 5G 移动网络的杀手级应用，但目前由于移动终端计算能力低、移动传输能力不稳定、移动网络服务时延大等原因，难以满足移动 AR/VR 的信息处理与传输需求，用户体验差强人意。未来 5G 移动网络将是一个异构通信与多级计算融合的网络，协同通信与计算、存储资源，有可能大幅提升网络性能，满足移动 AR/VR 的信息处理与传输需求。面向通信与计算融合，未来 5G 的移动 AR/VR 信息处理与传输应在基于多级计算的移动 AR/VR 海量信息处理、融合计算的移动 AR/VR 智能传输机制、保障 AR/VR 时延的移动网络异质资源协同与优化等方面展开研究，深入探讨其中的机理，抽象出理论模型，并提出高效的机制，大幅改善移动 AR/VR 的用户体验，促进 5G 发展。

参考文献：

[1] 3GPP. System architecture for the 5G system, version 1.2.0: TS23.501[S]. 2017.

[2] ETSI. Mobile edge computing (MEC) ETSI Industry Specification Group (ISG) version 1.1.1[R]. 2016.

[3] 德意志银行. 了解关于 VR 的一切[R]. 2016.
Deutsche Bank. Learn all about VR[R]. 2016.

[4] 齐彦丽, 周一青, 刘玲, 等. 融合 MEC 的未来 5G 移动通信网络[J]. 计算机研究与发展, 2018, 55(3): 478-486.
QI Y L, ZHOU Y Q, LIU L, et al. MEC coordinated future 5G mobile wireless networks[J]. Journal of Computer Research and Development, 2018, 55(3): 478-486.

[5] 周一青, 李国杰. 未来移动通信系统中的通信与计算融合[J]. 电信科学, 2018, 34(3): 1-7.
ZHOU Y Q, LI G J. Convergence of communication and computing infuture mobile communication systems[J]. Telecommunications Science, 2018, 34(3): 1-7.

[6] DAME A, PRISACARIU V A, REN C, et al. Dense reconstruction using 3D object shape priors[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. Washington DC: IEEE Computer Society, 2013: 1288-1295.

[7] PRADEEP V, RHEMANN C, IZADI S, et al. MonoFusion: real-time 3D reconstruction of small scenes with a single web camera[C]//IEEE International Symposium on Mixed and Augmented Reality, Oct 1-4, 2013, Adelaide, Australia. Washington DC: IEEE Computer Society, 2013: 83-88.

[8] LAN G, LUO Z, HAO, et al. Development of a virtual reality teleconference system using distributed depth sensor[C]//IEEE International Conference on Computer and Communications, December 13-16, 2017, Chengdu, China. Washington DC: IEEE Computer Society, 2017: 975-978.

[9] KAPUSTA P, MAJCHROWICZ M, SANKOWSKI D, et al. Acceleration of image reconstruction in 3D electrical capacitance tomography in heterogeneous, multi-GPU system using sparse matrix computations and finite element method[C]//The Federated Conference on Computer Science and Information Systems, September 11-14, 2016, Gdansk, Poland. Washington DC: IEEE Computer Society, 2016: 679-683.

[10] QIAN M, WANG Y, ZHOU Y, et al. A super base station based centralized network architecture for 5G mobile communication systems[J]. Digital Communications and Networks, 2015, 1(2): 152-159.

[11] ZHOU Y, LIU H, PAN Z, et al. Energy efficient two-stage cooperative multicast based on device to device transmissions: effect of user density[J]. IEEE Transactions on Vehicular Technology, 2016, 65(9): 7297-7307.

[12] SUN Q, TIAN L, ZHOU Y, et al. Energy efficient incentive resource allocation in D2D cooperative communications[C]//IEEE International Conference on Communications (ICC), June 8-12, 2015, London, UK. Washington DC: IEEE Computer Society, 2015: 2632-2637.

[13] LOCHER A, PERDOCH M, RIEMENSCHNEIDER H, et al. Mobile phone and cloud-a dream team for 3D reconstruction[J]. IEEE Applications of Computer Vision, 2016: 1-8.

[14] MASIP-BRUIN X, MARÍN-TORDERA E, TASHAKOR G, et al. Foggy clouds and cloudy fogs: a real need for coordinated management of fog-to-cloud computing systems[J]. IEEE Wireless Communications, 2016, 23(5): 120-128.

[15] ZHENG K, MENG H, CHATZIMISIOS P, et al. An SMDP-based resource allocation in vehicular cloud computing systems[J]. IEEE Transactions on Industrial Electronics, 2015, 62(12): 7920-7928.

[16] BONOMI F, MILIT RO, ZHU J, et al. Fog computing and its role in the internet of things[C]//ACM Workshop on Mobile Cloud Computing, August 17, 2012, Helsinki, Finland. New York: ACM Press, 2012: 13-16.

[17] LIU Y, LEE M J, ZHENG Y, et al. Adaptive multi-resource allocation for cloudlet-based mobile cloud computing system[J]. IEEE Transactions on Mobile Computing, 2016, 15(10): 2398-2410.

[18] ETSI. Mobile edge computing (MEC): framework and reference architecture, version 1.1.1[R]. 2016.

[19] HOU X, LU Y, DEY S, et al. Wireless VR/AR with edge/cloud computing[C]//International Conference on Computer Communication and Networks, September 18, 2017, Vancouver, Canada. Washington DC: IEEE Computer Society, 2017: 1-8.

[20] TANSKANEN P, KOLEV K, MEIER L, et al. Live metric 3D reconstruction on mobile phones[C]//IEEE International Con-

ference on Computer Vision, March 3, 2013, Sydney, Australia. Washington DC: IEEE Computer Society, 2013: 65-72.

[21] KOLEV K, TANSKANEN P, SPECIALE P, et al. Turning mobile phones into 3D scanners[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, USA. Washington DC: IEEE Computer Society, 2014: 3946-3953.

[22] ONDR´UˇSKA P, KOHLI P, IZADI S, et al. Mobile fusion: real-time volumetric surface reconstruction and dense tracking on mobile phones[J]. IEEE Transactions on Visualization & Computer Graphics, 2015, 21(11): 1251-1258.

[23] MUR-ARTAL R, MONTIEL J M M, TARD´OS J D, et al. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[24] LAI Z, HU Y C, CUI Y, et al. Furion: engineering high-quality immersive virtual reality on today's mobile devices[C]//ACM 23rd Annual International Conference on Mobile Computing and Networking, October 16-20, 2017, Snowbird, Utah, USA. New York: ACM Press, 2017: 409-421.

[25] LIU L, ZHOU Y, TIAN L, et al. CPC-based backward compatible network access for LTE cognitive radio cellular networks[J]. IEEE Communication Magazine, 2015, 53(7): 93-99.

[26] SIMON N. Cognitive radio: brain-empowered wireless communications[J]. IEEE Journal on Selected Areas in Communications, 2005, 23(2): 201-220.

[27] YIN W S, REN P Y, DU Q H, et al. Delay and throughput oriented continuous spectrum sensing schemes in cognitive radio networks[J]. IEEE Transactions on Wireless Communications, 2012, 11(6): 2148-2159.

[28] QUAN Z, CUI S, SAYED A H, et al. Optimal linear cooperation for spectrum sensing in cognitive radio[C]//IEEE Military Communications Conference, Oct 29-31, 2007, Orlando, FL, USA. Washington DC: IEEE Computer Society, 2007: 1-6.

[29] ZHAO Y, MORALES L, GAEDDERT J, et al. Applying radio environment maps to cognitive wireless regional area networks[C]//IEEE 2nd International Symposium on New Frontiers in Dynamic Spectrum Access Networks, Apr 17-20, 2007, Dublin, Ireland. Washington DC: IEEE Computer Society, 2007: 115-118.

[30] PEREZ-ROMERO J, ZALONIS A, BOUKHATEM L, et al. On the use of radio environment maps for interference management in heterogeneous networks[J]. IEEE Communications Magazine, 2015, 53(8): 184-191.

[31] WANG Y, PEDERSEN K I, SORENSEN T B, MOGENSEN P E, et al. Utility maximization in LTE-advanced systems with carrier aggregation[C]//IEEE Vehicular Technology Conference, Sept 5-8, 2011, Yokohama, Japan. Washington DC: IEEE Computer Society, 2011: 1-5.

[32] LIAO H S, CHEN P Y, CHEN W T, et al. An efficient downlink radio resource allocation with carrier aggregation in LTE-advanced networks[J]. IEEE Transactions on Mobile Computing, 2014, 13(13): 2229-2239.

[33] SUNDARESAN K, RANGARAJAN S. Energy efficient carrier aggregation algorithms for next generation cellular networks[C]//IEEE 21st International Conference on Network Protocols, Oct 7-10, 2013, Goettingen, Germany. Washington DC: IEEE Computer Society, 2013: 1-10.

[34] CHUNG Y L, JANG L J, TSAI Z, et al. An efficient downlink packet scheduling algorithm in LTE-advanced systems with carrier aggregation[C]//IEEE Consumer Communications and Networking Conference, Jan 10-13, 2011, Las Vegas, NV, USA. Washington DC: IEEE Computer Society, 2011: 632-636.

[35] WANG Y, PEDSRSEN K I, SORENSEN T B, MOGENSEN P E, et al. Carrier load balancing and packet scheduling for multi-carrier systems[J]. IEEE Transactions on Wireless Communications, 2010, 9(5): 1780-1789.

[36] SCHWARZ H, MARPE D, WIEGAND T, et al. Overview of the scalable video coding extension of the H.264/AVC standard[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2007, 17(9): 1103-1120.

[37] BOYCE J M, Ye Y, CHEN J, et al. Overview of SHVC: scalable extensions of the high efficiency video coding standard[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2016, 26(1): 20-34.

[38] SULLIVAN G J, BOYCE J M, CHEN Y, et al. Extensions of high efficiency video coding (HEVC)[J]. IEEE Journal of Selected Topics in Signal Processing, 2013, 7(6): 1001-1016.

[39] ZHAO Y, CHEN Z, ZHU C, et al. Binocular just-noticeable-difference model for stereoscopic images[J]. IEEE Signal Processing Letters, 2010, 18(1): 19-22.

[40] 周俊明, 郁梅, 蒋刚毅, 等. 面向 ROI 编码的立体图像比特分配策略分析[J]. 高技术通讯, 2011, 21(10): 1048-1055.
ZHOU J M, YU M, JIANG G Y, et al. Analysis of bit allocation for ROI based coding of stereoscopic images[J]. Chinese High Technology Letters, 2011, 21(10): 1048-1055.

[41] SHAO F, JIANG G, YU M, et al. A novel rate control technique for asymmetric-quality stereoscopic video[J]. IEEE Transactions on Consumer Electronics, 2012, 57(4): 1823-1829.

[42] WEI H, ZHOU X, ZHOU W, et al. Visual saliency based perceptual video coding in HEVC[C]//IEEE International Symposium on Circuits and Systems, May 22-25, 2016, Florence, Italy. Piscataway: IEEE Press, 2016: 2547-2550.

[43] KIM J, BAE S H, KIM M, et al. An HEVC-compliant perceptual video coding scheme based on JND models for variable block-sized transform kernels[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2015, 25(11): 1786-1800.

[44] CHEN M, SAAD W, YIN C, et al. Virtual reality over wireless networks: quality-of-service model and learning-based resource management[J]. IEEE Transactions on Communications, 2018 (99): 1.

[45] AHN J, KIM Y Y, KIM R Y, et al. Delay oriented VR mode WLAN for efficient wireless multi-user virtual reality device[C]//IEEE International Conference on Consumer Electronics, Jan 8-10, 2017, Las Vegas, NV, USA. Washington DC: IEEE Computer Society, 2017: 122-123.

[46] LAI C F, CHAO H C, LAI Y X, et al. Cloud-assisted real-time transrating for HTTP live streaming[J]. IEEE Wireless Communications, 2013, 20(3): 62-70.

[47] WANG F, LIU J, CHEN M, et al. Calms: cloud-assisted live media streaming for globalized demands with time/region diversities[C]//IEEE INFOCOM, Mar 25-30, 2012, Orlando, FL, USA. Washington DC: IEEE Computer Society, 2012: 199-207.

[48] WANG F, LIU J, CHEN M, et al. Migration towards cloud-assisted live media streaming[J]. IEEE/ACM Transactions on Networking, 2016, 24(1): 272-282.

[49] HE J, XUE Z, WU D, et al. CBM: online strategies on cost-aware buffer management for mobile video streaming[J]. IEEE Transactions on Multimedia, 2014, 16(1): 242-252.

[50] ZHENG Y, WU D, KE Y, et al. Online cloud transcoding and distribution for crowd sourced live game video streaming[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2016: 1.

[51] TANG J, QUEK T Q S. The role of cloud computing in content-centric mobile networking[J]. IEEE Communications Magazine, 2016, 54(8): 52-59.

[52] DENG M, TIAN H, FAN B. Fine-granularity based application offloading policy in small cell cloud-enhanced networks[C]//IEEE International Conference on Communications Workshops, May 23-27, 2016, Kuala Lumpur, Malaysia. Washington DC: IEEE Computer Society, 2016: 638-643.

[53] CHEN X, JIAO L, LI W, et al. Efficient multi-user computation offloading for mobile-edge cloud computing[J]. IEEE/ACM Transactions on Networking, 2016, 24(5): 2795-2808.

[54] CAO S, TAO X, HOU Y. An energy-optimal offloading algorithm of mobile computing based on HetNets[C]//International Conference on Connected Vehicles and Expo, Oct 19-23, 2015, Shenzhen, China. Washington DC: IEEE Computer Society, 2015: 254-258.

[55] ZHAO Y, ZHOU S, ZHAO T, et al. Energy-efficient task offloading for multiuser mobile cloud computing[C]//IEEE/CIC International Conference on Communications in China, Oct 14-17, 2016, Shenzhen, China. Washington DC: IEEE Computer Society, 2016: 1-5.

[56] VONDRA M, BECVAR Z. QoS-ensuring distribution of computation load among cloud-enabled small cells[C]//IEEE International Conference on Cloud Networking, Oct 8-10, 2014, Luxembourg. Washington DC: IEEE Computer Society, 2014: 197-203.

[57] MUNOZ O, PASCUAL-ISERTE A, VIDAL J. Joint allocation of radio and computational resources in wireless application offloading[C]//IEEE Future Network and Mobile Summit, Jul 3-5, 2013, Lisboa, Portugal. Washington DC: IEEE Computer Society, 2013: 1-10.

[58] MAO Y, ZHANG J, SONG S H, et al. Power-delay trade-off in multi-user mobile-edge computing systems[C]//IEEE Global Communications Conference, Dec 4-8, 2016, Washington DC, USA. Washington DC: IEEE Computer Society, 2016: 1-6.

[59] LIU J, MAO Y, ZHANG J, et al. Delay-optimal computation task scheduling for mobile-edge computing systems[C]//IEEE International Symposium on Information Theory, Jul 10-15, 2016, Barcelona, Spain. Washington DC: IEEE Computer Society, 2016: 1451-1455.

[60] MAO 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.

[61] SATYANARAYANAN M, BAHL P, DAVIES N. The case for VM-based cloudlets in mobile computing[J]. IEEE Pervasive Computing, 2009, 8(4): 14-23.

[62] WANG K, SHEN M, CHO J, et al. MobiScud: a fast moving personal cloud in the mobile network[C]//ACM Workshop on All Things Cellular: Operations, Applications and Challenges, Aug 17, 2015, New York, USA. New York: ACM Press, 2015: 19-24.

[63] BECVAR Z, PLACHY J, MACH P. Path selection using handover in mobile networks with cloud-enabled small cells[C]//IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication, Aug 30-Sept 2, 2015, Washington DC, USA. Washington DC: IEEE Computer Society, 2015: 1480-1485.

[64] RAVI A, PEDDOJU S K. Handoff strategy for improving energy efficiency and cloud service availability for mobile devices[J]. Wireless Personal Communications, 2015, 81(1): 101-132.

[65] CLARK C, FRASER K, HAND S, et al. Live migration of virtual machines[C]//2nd Conference on Symposium on Networked Systems Design and Implementation, May 2-4, 2005, Lombard, IL, USA. New York: ACM Press, 2005: 273-286.

[66] BELLAVISTA P, ZANNI A, SOLIMANDO M. A migration-enhanced edge computing support for mobile devices in hostile environments[C]//IEEE Wireless Communications and Mobile Computing Conference, June 26-30, 2017, Valencia, Spain. Washington DC: IEEE Computer Society, 2017: 957-962.

[67] WANG S, URGAONKAR R, CHAN K, et al. Dynamic service placement for mobile micro-clouds with predicted future costs[C]//IEEE International Conference on Communications, June 8-12, 2015, London, UK. Washington DC: IEEE Computer Society, 2015: 5504-5510.

[68] PLACHY J, BECVAR Z, STRINATI E C. Dynamic resource allocation exploiting mobility prediction in mobile edge computing[C]//IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications, Sept 4-8, 2016, Valencia, Spain. Washington DC: IEEE Computer Society, 2016: 1-6.

[69] FANG W, ZHENG L, DENG H, et al. Real-time motion tracking for mobile augmented/virtual reality using adaptive visual-inertial fusion[J]. Sensors, 2017, 17(5): 1-22.

[70] KIM H, LEUTENEGGER S, DAVISON A J. Real-time 3D reconstruction and 6-DoF tracking with an event camera[C]//European Conference on Computer Vision, October 8-16, 2016, Amsterdam, Netherlands. New York: Springer International Publishing, 2016: 349-364.

[71] DOBBINS M K, RONDOT P, SCHWARTZ K, et al. Providing a collaborative immersive environment using a spherical camera and motion capture: US 8217995 B2[P]. 2012-07-10.

[72] AHLEHAGH H, DEY S. Hierarchical video caching in wireless cloud: approaches and algorithms[C]//IEEE International Conference on Communications, June 10-15, 2012, Ottawa, ON, Canada. Washington DC: IEEE Computer Society, 2012: 7082-7087.

[73] WANG X, CHEN M, TALEB T, et al. Cache in the air: exploiting content caching and delivery techniques for 5G systems[J]. IEEE Communications Magazine, 52(2): 131-139.

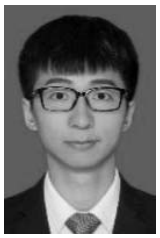
[74] MOLISCH A F, CAIRE G, OTT D, et al. Caching eliminates the wireless bottleneck in video aware wireless networks[J]. Advances in Electrical Engineering, 2014(9): 74-80.

[75] LI W, OTEAFY S M A, HASSANEIN H S. Dynamic adaptive

streaming over popularity-driven caching in information-centric networks[C]//IEEE International Conference on Communications, June 8-12, 2015, London, UK. Washington DC: IEEE Computer Society, 2015: 5747-5752.
[76] LIU Y, GEURTS J, POINT J C, et al. Dynamic adaptive streaming over CCN: a caching and overhead analysis[C]//IEEE International Conference on Communications, June 9-13, 2013, Budapest, Hungary. Washington DC: IEEE Computer Society, 2013: 2222-2226.
[77] LIANG K, HAO J, ZIMMERMANN R, et al. Integrated prefetching and caching for adaptive video streaming over HTTP: an online approach[C]//ACM Multimedia Systems Conference, March 18-20, 2015, Portland, OR, USA. New York: ACM Press, 2015: 142-152.
[78] GRANDL R, SU K, WESTPHAL C. On the interaction of adaptive video streaming with content-centric networking[J]. Journal of Sports Sciences, 2013, 23(9): 977-989.
[79] FUENTE Y S D L, SCHIERL T, HELLGE C, et al. iDASH: improved dynamic adaptive streaming over HTTP using scalable video coding[C]//ACM SIGMM Conference on Multimedia Systems, Feb 23-25, 2011, San Jose, CA, USA. New York: ACM Press, 2011: 257-264.
[80] TRAN T X, POMPILI D. Octopus: a cooperative hierarchical caching strategy for radio access networks[C]//IEEE International Conference on Mobile Ad Hoc and Sensor Systems, Oct 10-13, 2016, Brasilia, Brazil. Washington DC: IEEE Computer Society, 2016: 154-162.
[81] ZHANG Z, LIU D, YUAN Y. Layered hierarchical caching for SVC-based HTTP adaptive streaming over C-RAN[C]//IEEE Wireless Communications and Networking Conference, March 19-22, 2017, San Francisco, CA, USA. Washington DC: IEEE Computer Society, 2017:1-6.

[作者简介]

周一青（1975−），女，中国科学院大学岗位教授，中国科学院计算技术研究所“百人计划”研究员、博士生导师，无线通信技术研究中心副主任，移动计算与新型终端北京市重点实验室研究员，主要研究方向为移动通信、通信与计算融合等。



孙布勒（1992−），男，中国科学院计算技术研究所博士生，主要研究方向为通信与计算融合、大规模多天线、毫米波通信等。

齐彦丽（1991−），女，中国科学院计算技术研究所博士生，主要研究方向为移动边缘计算、通信与计算融合、无线资源管理等。

彭燕（1993−），女，中国科学院计算技术研究所博士生，主要研究方向为通信与计算融合、超密集网络等。

刘玲（1990−），女，中国科学院计算技术研究所助理研究员，主要研究方向为 5G 无线通信、通信与计算融合、超密集网络、无线资源管理等。

张志龙（1985−），男，北京邮电大学讲师，主要从事多媒体通信、无线视频传输策略优化等方面的研究工作。

刘奕彤（1982−），女，北京邮电大学讲师，主要从事 VR 视频和三维模型的无线传输等方面的研究工作。

刘丹谱（1972−），女，北京邮电大学国际学院电信工程及管理系主任、博士生导师，IEEE 高级会员，中国通信学会和中国电子学会高级会员，一直致力于无线视频传输策略优化领域的基础理论与关键技术研究。

李兆歆（1983−），男，中国科学院计算技术研究所助理研究员，主要从事三维重建领域的研究工作。

田霖（1980−），女，中国科学院计算技术研究所副研究员，主要研究方向为绿色无线通信系统与无线资源管理技术。

专题：5G

5G 先进技术研究进展

林泓池，孙文彬，郭继冲，麻津铭，周永康，于启月，孟维晓
（哈尔滨工业大学，黑龙江 哈尔滨 150001）

摘　要：5G 移动通信技术的标准制定正在如火如荼地进行中，相比前几代移动通信技术，5G 面临着更复杂的业务需求、更极致的用户体验和更密集的网络架构。而且 5G 需要在大数据和人工智能的时代到来之前，做好万物互联的通信基础。对目前 5G 的一些热点技术进行了简单的介绍，首先给出了 5G 的技术愿景和需求目标，然后对大规模天线技术、超密集组网技术和非正交多址技术 3 个热点技术的研究进展情况进行了简单的阐述。

关键词：5G；大规模天线阵列；超密集组网；非正交多址接入
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018238

Research progress of 5G advanced technologies

LIN Hongchi, SUN Wenbin, GUO Jichong, MA Jinming,
ZHOU Yongkang, YU Qiyue, MENG Weixiao
Harbin Institute of Technology, Harbin 150001, China

Abstract: The standards of the 5th generation communication technology were picking up. Compared with previous generations' communication technologies, 5G will face more complex business requirements, more extreme user experience and more dense network architecture. Moreover, 5G needs to lay a good communication foundation for IoT before the arrival of the era of big data and artificial intelligence. Some hot technologies of 5G were briefly introduced. Firstly, the technical vision and demand target of 5G was given, and then the research progress of three hot technologies of massive MIMO, ultra dense network and non-orthogonal multi-access technology was expounded.

Key words: 5G, massive MIMO, UDN, NOMA

1　引言

随着大数据和万物互联时代的即将到来，新任务和新需求对移动通信网络发展提出了极大的挑战，并积极地推进着 4G 时代向 5G 时代的转化。同时正值新兴技术的工业革命时期，5G 作为新一

收稿日期：2018−07−20；修回日期：2018−08−06
基金项目：国家自然科学基金资助项目（No.61728104）；黑龙江省自然科学基金重点项目（No.ZD2017013）
Foundation Items: The National Natural Science Foundation of China(No.61728104), The Natural Science Foundation Major Project of Heilongjiang Province(No.ZD017013)

代的移动通信技术，将为万物互联和人工智能的发展提供良好的通信保障。

一直以来新兴业务对通信质量的新需求和对通信环境的新愿景，促进着整个通信行业的不断发展和通信技术的不断更迭。与前几代移动通信技术相比，5G 的业务类型将更加丰富、复杂，导致在面对不同业务、不同需求、不同场景的业务无法统一时，5G 很难跟前几代的移动通信技术一样以某一种先进的关键技术作为基点形成解决方案来解决自己技术场景带来的困难和挑战。我国 IMT-2020（5G）推进组最开始根据 5G 需求和愿景，分析讨论了 5G 所需面临的挑战和 5G 未来的愿景以及愿景中所需要的适用关键技术，整理发布了 5G 概念白皮书[1]。而在 2015 年 6 月，国际电信联盟（International Telecommunication Union，ITU）正式以 IMT-2020 命名 5G，并根据业务需求特点及应用场景的不同，定义了 5G 的三大应用场景：增强移动宽带（eMBB）、大规模机器类通信（mMTC）和超可靠低时延类通信（uRLLC）[2]。2018 年 6 月 14 日，国际标准组织 3GPP 在美国举行全体会议，5G 移动通信技术标准的方案获得批准并发布，这标志着首个真正完整的国际 5G 标准正式出炉，5G 已完成第一阶段全功能标准化工作，进入了产业发展新阶段。5G 挑战与关键技术对应框图如图 1 所示。

与现有 4G 相比，用户数和用户需求的增加，导致网络密集化和业务多样化，同时 5G 还需要面对极致的性能要求，这些无疑对 5G 提出了极大的挑战。参考文献[3]给出的挑战具体包括用户体验速率扩大 10~100 倍、更高的频谱利用率、毫秒级的端到端时延、大规模连接数扩大 10~100 倍、更低的开销以及良好的用户体验。除此之外，为了实现与前几代移动通信网络的融合、过渡和可持续发展，5G 还需满足网络灵活部署和高效运营维护的要求。图 1 简单地描述了这些挑战以及 5G 独特的技术解决方案和相应的设计原则[4]。

5G 愿景的提出引发了一系列通信技术的变革和应用，在新频谱方面，有扩展频谱资源的毫米波通信技术；在通信模式方面，有大规模天线阵列技术、非正交多址接入技术和全双工通信技术；在组网模式方面，有超密集组网技术和 D2D

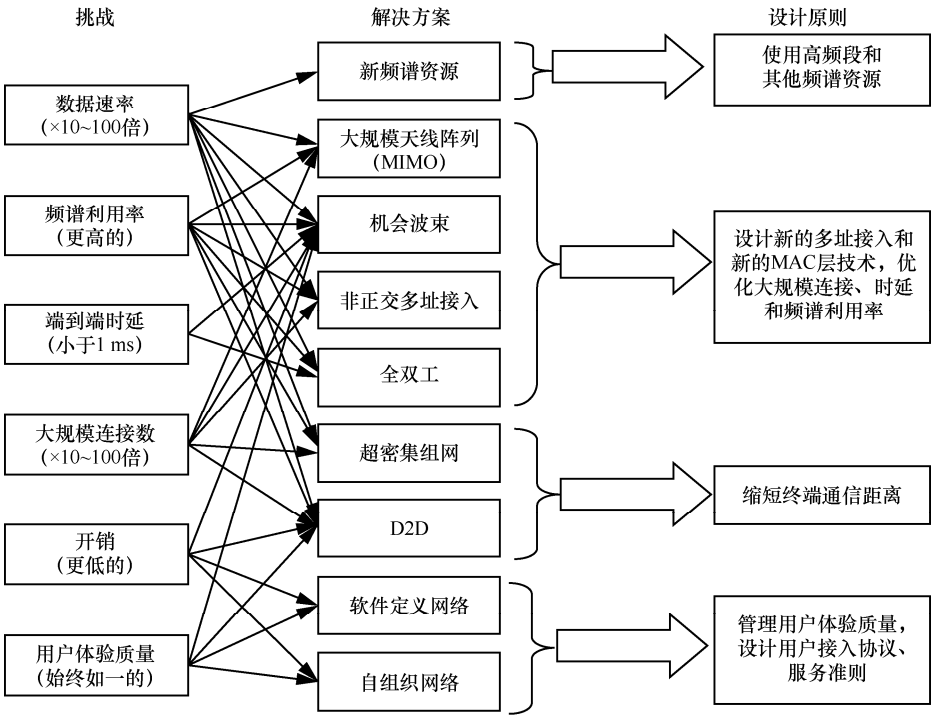


图 1　5G 挑战与关键技术对应框图[3]

通信技术；在网络架构方面，有软件定义网络架构和自组织网络架构；在编码方面，有适用于长码的 LDPC 码和适用于短码的 Polar 码。随着这些技术的研究发展和实践，5G 的系统将不断得到完善，并且 5G 的性能也将不断得到提升[3-6]。

本文将选取一些具有 5G 特色的先进技术进行介绍，并对其技术的研究现状与未来发展进行简单评述，内容包含 3 个方面：4G 的 MIMO 的改进技术——大规模天线阵列、5G 组网通信模式的愿景——超密集组网、5G 多址接入技术——非正交多址接入技术。

2 大规模天线阵列

如前所述，3GPP 规定了 5G 的三大应用场景，即 eMBB、mMTC 和 uRLLC。3 个场景的侧重点不同，对无线通信技术的要求也各不相同。而针对 eMBB 场景来说，要求量级在 10 Gbit/s 以上超高的信息传输速率。为了实现这么高速率的传输，可做的工作有：密集化节点、增加带宽、提升频谱效率[7]。而在 4G 中的 MIMO 技术，一般是 8 根天线的 MIMO，所以为了进一步提升频谱效率，5G 的目光放在了大规模 MIMO 技术。

大规模 MIMO 概念最早于 2010 年被 Marzetta 在参考文献[8]中详细阐述。在提出之初，就以超高的频谱效率和极其简单的收发机结构引起了广泛的关注。随着研究热度和深度的增加，基于大规模 MIMO 技术的相关理论研究越来越多，然而在实践工程中却出现了阻力。在大规模 MIMO 提出之初，采用全数字预编码方案对多用户干扰进行控制，如全数字迫零（zero-force，ZF）和全数字匹配滤波（matched filter，MF）[8]。当信道衰落服从独立同分布时，全数字 MF 预编码方案可以最低的计算复杂度获得最佳的性能。但是该方案要求射频链路数等于发射天线数，也就意味着基站需要配有成百上千的射频链路。这样的要求在实际工程中会导致成本太高，能耗太高是不能被满足的。起初面对此困境，研究者想到了全模拟预编码方案，其仅仅需要等于独立数据流数的射频链路数，容易在实际场景中实现。但是全模拟预编码方案却存在旁瓣干扰，致使性能较差。

针对 5G 的高速率低功耗的绿色通信的理念，提出了在全数字与全模拟预编码方案之间取了一个折中的混合预编码，是目前的大规模 MIMO 研究热点[9]。混合预编码方案由数字预编码和模拟预编码两部分组成，如图 2 所示，所需的射频链路数大于或等于独立数据流数但小于或等于 2 倍的独立数据流数[9]，能逼近全数字预编码方案的性能。

如图 3、图 4 所示在平均和速率方面上，混合预编码的性能可以取得接近于全数字预编码的性能，且明显优于全模拟预编码的性能，如图 3 所示。而在平均能耗方面上，混合预编码的性能不

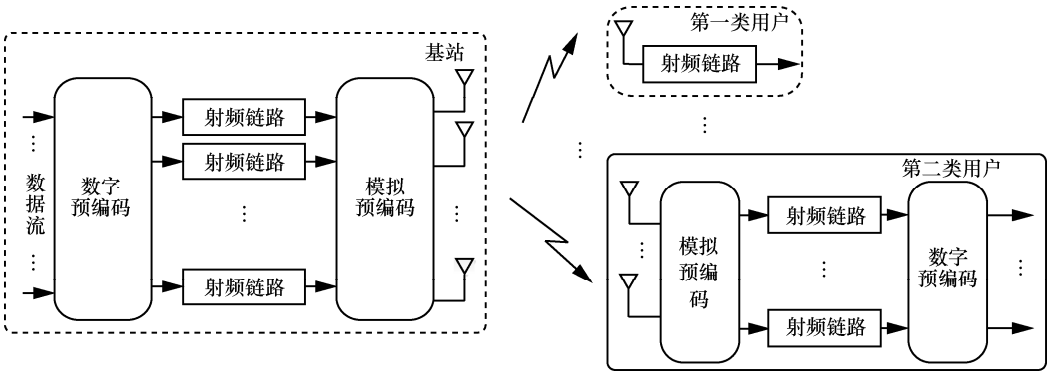


图 2 大规模 MIMO 混合预编码系统结构

如全模拟预编码的性能，但比全数字预编码的性能好，如图 4 所示。

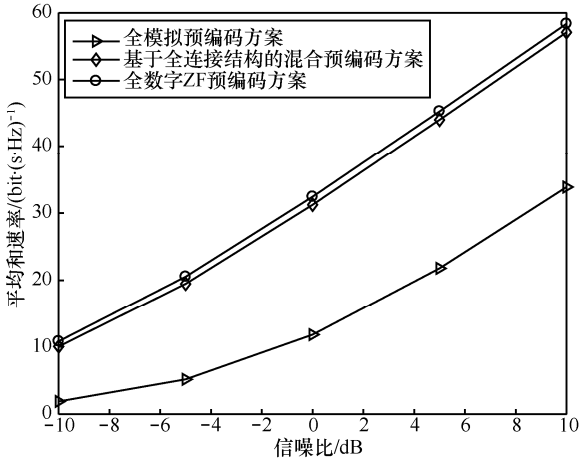


图 3　线性预编码的平均和速率

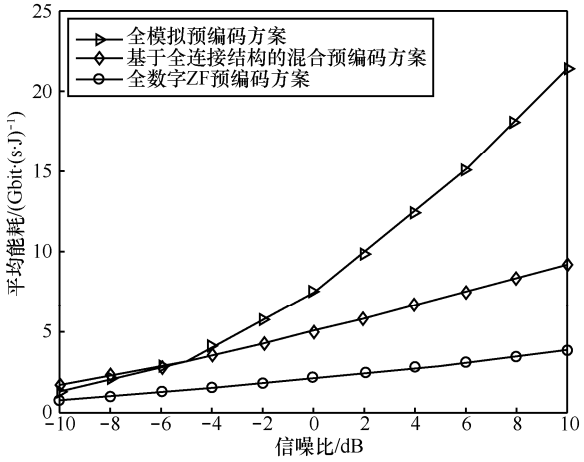


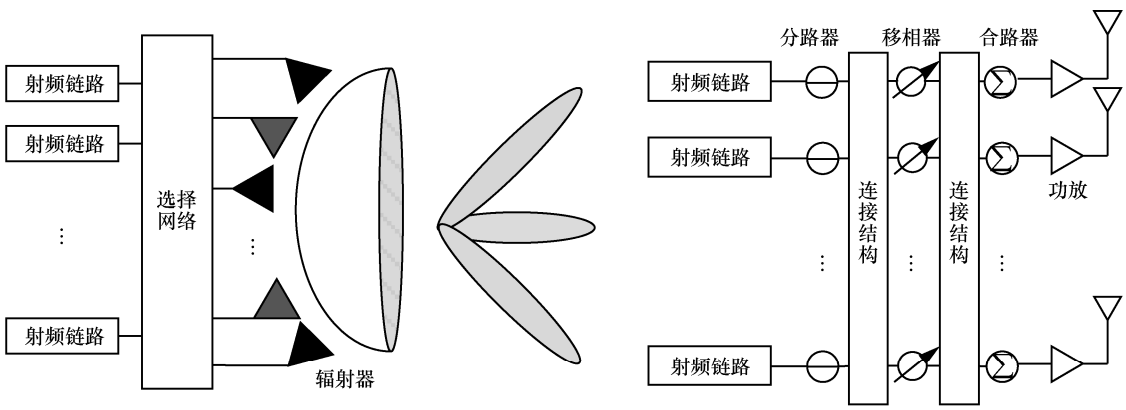
图 4　线性预编码的平均能效

为了更好地理解大规模 MIMO 在 5G 愿景中的应用，本文将围绕两个研究热点展开具体技术细节的介绍，即毫米波大规模 MIMO 波束成形技术和大规模 MIMO 机会波束成形技术。

2.1　毫米波大规模 MIMO

重视毫米波的研究和使用是 5G 的特点之一，这主要是由于毫米波具有丰富的频率资源。然而单纯的毫米波技术本身带有路径损耗大的特点，不适合远距离传输。所以将其与大规模 MIMO 结合，在弥补毫米波技术本身缺陷的同时，可以进一步提升系统的传输速率。毫米波大规模 MIMO 系统工作的频段在 30 ~300 GHz，对应的波长在 0.1~1 cm，十分有利于大规模天线阵列的部署，从而为系统提供较高的阵列增益。此外，由于高频电磁波的传播特性，毫米波大规模 MIMO 系统的信道模型具有很强的视距路径分量。这两个特点为毫米波大规模 MIMO 波束成形技术提供了基础，保证了其具有较低复杂度和较好性能的优点[10-11]。

根据阵列所用天线的不同，毫米波大规模 MIMO 波束成形技术可以分成两种：基于透镜天线阵列和基于射频天线阵列的波束成形技术，如图 5（a）所示。特殊地，经过精心设计的离散透镜天线阵列可以起到空域离散傅立叶变化的作用[12]。在毫米波传播环境中，由于毫米波本身易被散射体吸收，不易被散射体反射，导致散射簇的数量十分有限，波束空间信道呈现一种稀疏性，即有效波束的个数



(a) 透镜天线阵列[10]　　(b) 射频天线阵列

图 5　基于不同天线阵列的毫米波大规模 MIMO 波束成形技术

比较少。合适的波束选择算法可以在保证一定系统性能的情况下，使得需要工作的辐射器数量骤减，从而大大减少所需的射频链路数。因此，当前基于透镜天线阵列的波束成形技术的主要研究方向就是设计合适的波束选择算法，目前已经有大量的算法被提出[10-13]。

虽然透镜天线有着独特的特点，但是透镜天线生产效率低，不易构造，限制了透镜天线的使用。而普通的射频天线则没有这种缺点，且由其组成的大规模天线阵列可以拥有较小的旁瓣和后瓣，因此，基于射频天线阵列的波束成形技术引起了研究者的广泛关注[11,14]。改变图 5（b）中的连接结构，可以得到不同的射频天线阵列结构。目前，射频天线阵列的结构主要有 4 种，分别是全连接结构（fully-connected architecture）、子阵列结构（array-of-subarray architecture）、过载子阵列结构（overlapped subarray architecture，OSA）以及自适应子阵列结构（adaptive array-of-subarray architecture）。基于上述天线阵列结构，通信研究者已经提出了很多优秀的波束成形方案，其目标函数涉及系统有效性、可靠性、能效等[15,17]。

2.2 机会波束形成

传统的 MIMO 系统采用波束成形（预编码）技术来避免多个用户和数据流之间的干扰，从而提高系统性能。然而，波束成形技术最重要的前提条件是瞬时信道状况完全已知。因此，信道估计技术对于 MIMO 系统的性能尤为重要。随着用户数目和天线数目的增加，信道估计技术的复杂性呈指数倍增加。并且在现实的物理环境中，信道估计技术的复杂性是极其高的，跟踪信道的瞬时状态从而获得完美的信道信息是一件不可能完成的任务。所以对于 5G 如果想要用大规模的天线来提高系统性能来说，在无法获得完美信道信息的状况下，传统波束成形的系统性能会有大幅度的下降。而机会波束成形可以在无法获得完美信道信息的状况下，依然保持一定的系统性能。并在用户密集的情形下，会有用户分集和系统简单能耗低的优点。所以机会波束成形可以作为 5G 物联网的备选技术之一。

机会波束成形系统模型如图 6 所示，即通过将多天线技术和多用户机会分配技术相结合，机会波束成形技术可以在低反馈链路的情况下获得多用户分集增益。在机会波束成形系统中，每一个传输天线 k 上都乘上一个随机的矩阵 W_k。随机矩阵 W_k 和真实物理信道矩阵 H_k 的内积被定义为等效信道 h^{equ}，其接收信号表达式如式（1）所示：

$$y_k = H_k W_k x_k + z_k = \sum_{n=1}^{N_T} h_n w_n x_k + z_k = h^{\text{equ}} x_k + z_k \quad (1)$$

其中，y_k 表示接收信号，H_k 为信道矩阵，W_k 为随机波束成形矩阵，x_k 为用户数据，z_k 为加性高斯白噪声，h_n 表示第 n 个天线的信道系数，w_n 表示第 n 个天线的随机成形系数。

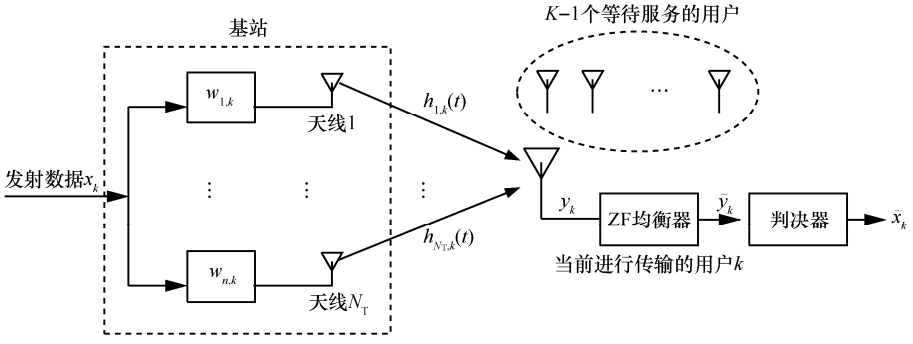


图 6　机会波束成形系统模型

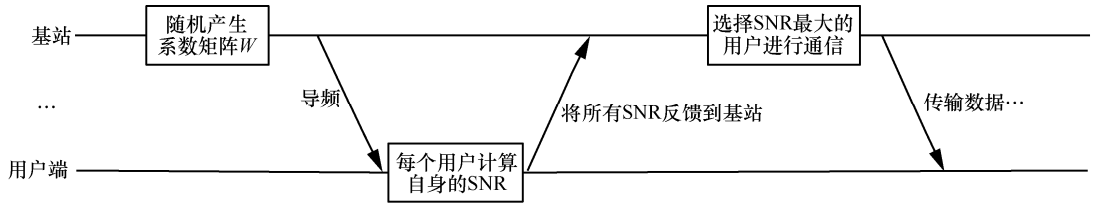


图 7　机会波束成形传输结构

不同于传统的波束成形系统，等效信道决定了接收的信噪比和系统性能。在接收端，每个用户测量自身的信噪比来衡量等效信道的质量，并且将所测量的信噪比反馈给基站。其反馈的信噪比为：

$$\mathrm{SNR}_k = \frac{\left|h_k^{\mathrm{equ}}\right|^2}{2\sigma_{z_k}^2} \tag{2}$$

其中，$\sigma_{z_k}^2$ 为加性高斯白噪声的方差。这里对发射信号的功率进行了归一化。

基站根据反馈的信噪比选择最大信噪比的用户来传输数据，从而实现最大的系统容量和最小的误码率。其传输结构如图 7 所示。

与其他波束成形技术相比较，机会波束成形的波束成形矩阵设计比较简单，并且当真实的物理信道变化比较缓慢的时候，机会波束成形技术可以通过提高随机矩阵的变化速率来改善系统性能。同时由于机会波束成形技术并不需要较多的信道和用户信息，因此机会波束成形技术可以很容易地与其他多用户复用技术相结合，从而同时获得多用户分集和复用增益，例如时分多址—机会波束成形系统、频分多址—机会波束成形系统以及非正交多址—机会波束成形系统。

图 8 给出了机会波束成形技术与重复编码（repetition coding，RC）、空时编码（space-time block coding，STBC）在瑞利信道以及莱斯信道的误码性能的比较，与其他分集技术比较，可以看出机会波束成形技术有良好的误码性能。图 9 给出了机会波束成形技术与矢量量化（vector quantization，VQ）、格拉斯曼子空间包（Grassmannian subspace packing，GSP）技术[18]以及基因算法（genetic algorithm，GA）[19]在低反馈情况下的误码率比较，可以看出机会波束成形技术的误码性能最好。因此可知机会波束成形技术可以实现最优的系统性能[20]。

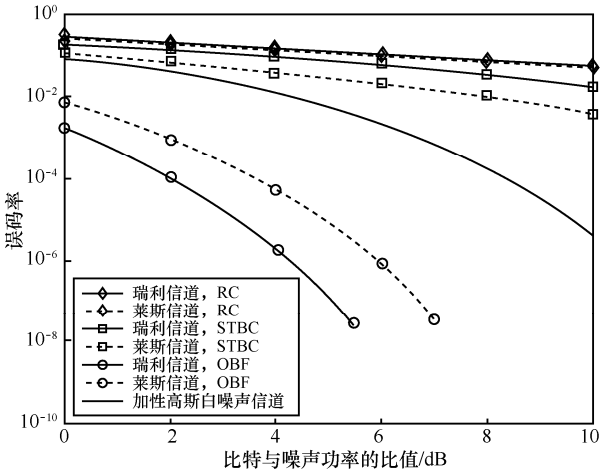


图 8　机会波束成形与其他分集技术的误码率比较[20]

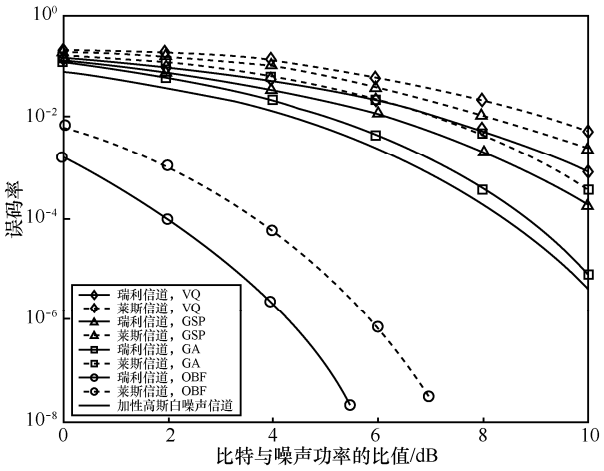


图 9　机会波束成形与其他波束成形技术的误码率比较[20]

然而对于机会波束成形技术而言，由于每个用户需要在被服务前加入等待队列，同时基站总是选择反馈信噪比最大的用户进行服务，用户之

间的公平性和用户服务的实时性很难得到保障，因此未来可以研究机会波束成形系统中的用户公平性问题，并提出一个基于用户实时性服务的改进机会波束成形系统。而且对于用户数目较少的情况，机会波束成形技术与传统的波束成形技术相比，并没有较大的性能优势。因此针对低用户数目情况下，如何提高机会波束成形系统性能也是未来研究机会波束成形技术的一个研究热点。

随着研究的不断深入，大规模 MIMO 作为 5G 应用中的核心技术之一被寄予了厚望，与现有技术结合的大规模 MIMO 也将会在 5G 应用上大展宏图。

3 超密集组网技术

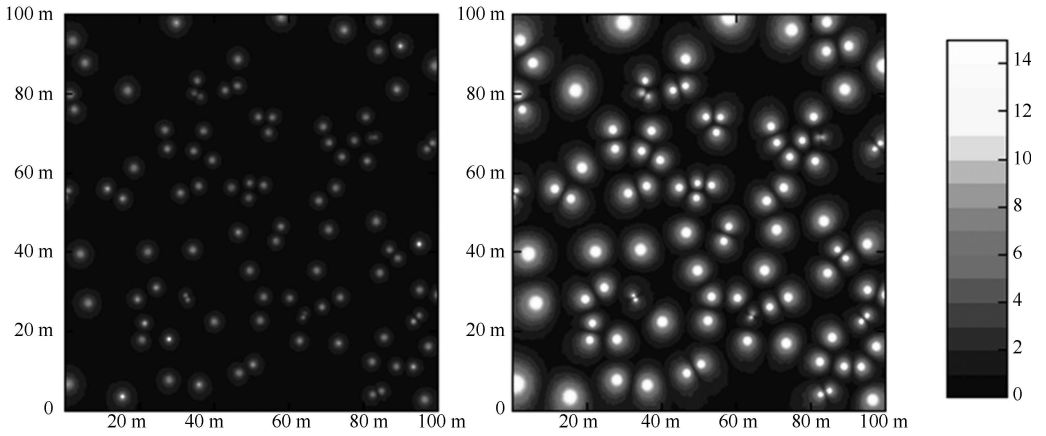
与 4G 移动通信技术相比，5G 网络的容量需求将有 1 000 倍的增长，届时，通信终端无处不在，并且在大型的热点区域，如商场、露天展台等地方上将存在着大量的设备需要连接。与此同时，不同的移动终端会带来不同的业务需求，这也就导致了业务需求的多样化。这使得移动蜂窝小区里的用户越来越密集，简单的和单一的通信网络架构不足以支撑非常密集和业务多样化的蜂窝小区的用户需求。因此提出了一个异构网络架构的超密集组网（ultra dense network，UDN），将用于满足区域面积内超高的容量需求（即热点问题），为移动终端提供无缝的网络切换，让用户在任何时间、任何地点都能拥有超高速的上网和通话体验。超密集组网通过在宏基站（macro base station，MBS）的热点区域放置微基站（small base station，SBS）形成微小区提供了更高的频谱自由度，有效地提高了系统的单位面积谱效率，从而提升了系统的性能[21]。

对宏小区中宏基站的部署一般多采用固定的格形部署，而微小区的微基站部署受周边环境和当地人流的影响。不仅如此，微小区的网络也呈现自组织性，不同层的网络架构之间也可能需要协作和补偿，种种原因导致了对微小区网络的建模不能单单采用传统的基于格点的基站部署方式[22]。同时由于用户在小区中呈现的分布不同，比如在某些热点区域（如展会中心区、景观区），用户分布较集中，呈现从中间向四周蔓延的趋势，而在其他区域，用户不会有明显的集聚效应，所以微小区的网络部署是超密集组网研究的难点和热点。

超密集组网在带来好处的同时，也带来了新的挑战，密集的网络使得基站之间的距离更近，随之带来的小基站之间干扰（inter cell interference）问题越发明显。不仅如此，网络的密集化也导致了小区中的干扰管理算法变得越发复杂。密集的组网也导致了很高的能耗[23]，SBS 能提供的功率也不会像宏基站提供的功率那么强，有一定的约束，如何将有限的功率很好地利用，使服务的用户更多，提供的容量更大，基站的耗能更小，也是现在亟需解决的热点问题。因此需要有新的干扰管理技术，充分协调现有的或者将要部署的 SBS 之间的关系，从而消除或减小基站之间的干扰，提升用户服务质量（quality of service，QoS），提升网络的单位面积频谱利用率，达到蜂窝小区总吞吐量提升的目的。与此同时，在超密集组网场景下，基站之间距离较近，基站之间的协作更便利，因此也可以使用多点联合传输（coordinated multipoint，CoMP） 技术进一步提升网络的性能。

图 10 显示了当路径损耗系数 $\alpha=2$ 和 $\alpha=4$ 的情况下，网络 100×100 中各个位置的遍历容量。可以看出 α 越大，随着用户距离基站的距离变大，用户的接收功率下降越快。导致用户接收到的有用信号的功率是降低的，但与此同时，由于 α 增大也使得用户接收到干扰基站的信号变小，因此，用户如何根据自己的路径损耗系数选取基站为自己服务，从而获得最大的遍历容量仍需要详尽的研究。

在超密集组网场景下，由于系统中的用户较



（a）α=2　　（b）α=4

图 10　网络中每个位置的遍历容量示意图

多、容量需求大和业务种类多样化，因此需要网络能够提供更高的单位面积的频谱利用率，更灵活的自适应资源调度策略[24]。同时，随着通信逐步成为世界上最耗能的应用的发展趋势[25]，需要想办法在满足用户要求的同时，尽可能降低能耗，从而得到较高的能耗比。超密集组网（UDN）作为 5G 的关键技术之一，对其概念的描述、基站布置与基站协作的探索、网络的性能分析和具体干扰管理算法的研究，都将是未来超密集组网研究的重点和热点方向。

4　非正交多址接入技术

多址接入技术通常被看作现代移动通信系统的特征代表，从 TDMA、FDMA、CDMA 开始，到后来的 OFDM 和 MIMO。随着正交多址接入技术的研究发展，多址接入技术已经逐步成熟，但仅利用正交多址接入（orthogonal multiple access，OMA）技术远远不能满足 5G 大连接、超密集的愿景。所以 5G 还需要开发出新型的多址技术。所以本着特殊到一般的研究思想，很容易就想到正交多址接入技术需要过渡到非正交多址接入（non-orthogonal multiple access，NOMA）技术[26]。正交多址技术是通过在多个相互正交的资源块上去区分和服务不同用户的，而非正交多址技术则是在同一资源块对不同的用户提供服务，即把多用户的信息进行叠加传输，这样不仅可以有效提高系统频谱效率，还可以提高基站服务的用户数，比较适合 5G 的万物互联和高频谱的愿景。所以 5G 除了支持传统的 OFDMA 技术外，还支持 NOMA、SCMA 等多种新型多址技术[27-28]。

4.1　NOMA

单讲 NOMA，一般指的是基于功率域复用的新型多址接入技术，以在接收端进行串行干扰消除算法实现对接收到的叠加信号进行译码。通过仿真系统的验证，NOMA 与传统 OMA 相比，提高了吞吐量和频谱效率，然而串行干扰消除（successive interference cancellation，SIC）通过功率排序依次对用户进行译码，势必会造成对误码的累积效应[29]。

虽然 NOMA 的吞吐量相比 OMA 提升了，但其是以牺牲误码性能换取的，即 NOMA 相比 OMA 增加了功率复用组内用户间干扰。所以在组内功率分配时，怎样分配使得吞吐量和误码性能达到一个权衡；同时对于复用用户之间怎样分组聚类，也是一个难点问题，虽然可以证明信道差异越大的两用户复用比随机两用户复用，系统性能的吞吐量有更大的提升，但其误码性能却较差（信道状态好的用户对信道状态差的用户干扰太

大，导致不能正常译码）；而且不同功率复用组间的功率怎么分配，使得整体系统吞吐量达到最优，也是一个热点问题；而且是否存在一种会比串行干扰消除译码方式更好的、适合 NOMA 的译码方式也有待研究；还有 NOMA 是否可以实现分集和复用同时存在？这样系统既有分集增益，又会有用户增益。

NOMA 组内两用户的误码率情况如图 11 所示，可以看出用串行干扰消除算法用户 2 可以直接检测得到自己的信息，而用户 1 需要先检测出用户 2 的信息，然后在接收的信号中减去用户 2 的信息，再检测才能得到自己的信息。可以看出对于用户 1 来说误码累积未必是件坏事。这是因为检测用户 2 的误码累积后，再经过判决可能会导致正确译码。所以串行干扰消除算法的误码分析还是需要详尽研究的。

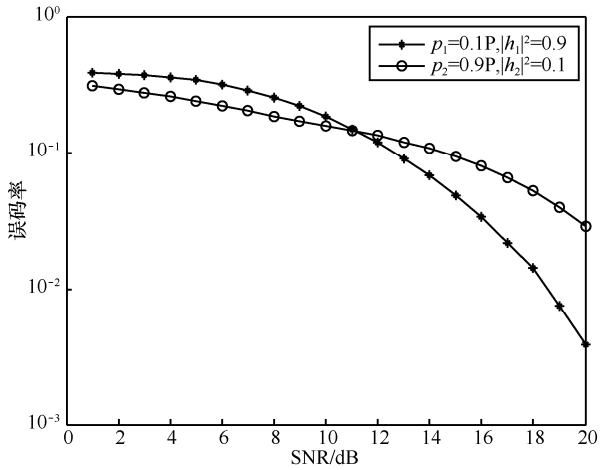


图 11　NOMA 组内两用户的误码率

（$p_1/p_2=1/9,|h_1|^2=0.9,|h_2|^2=0.1$）

除了以上 NOMA 本身的研究问题以外，NOMA 与其他技术的有机结合也是很有意思的课题，例如 OFDM 和 NOMA、MIMO 和 NOMA 等。

4.2　SCMA

2014 年，华为公司在参考文献[30]中提出了稀疏码多址接入（sparse code multiple access，SCMA）的概念，并认为 SCMA 可以看作低密度扩频（low density spreading，LDS）CDMA 的扩展和推广。在 SCMA 系统中，正交振幅调制（quadrature amplitude modulation，QAM）映射和扩频的过程融合在一起，形成一个 SCMA 码本，为 SCMA 编码过程带来成形增益[30]。像这样，每个传输层的二进制比特流都会根据码本直接映射成多维的复数码字，而用户通过不同的码字来实现共用信道，如图 12 所示，其中的每一列对应一组不同比特信息所映射的信号。

既然 SCMA 是码域的非正交多址接入技术，那么码本设计是 SCMA 中的一个核心问题，也是 SCMA 相关研究中比较具有挑战性的，SCMA 的码本可以形成一个多维的星座图，而多维星座图设计本身较为复杂[31-32]。在 SCMA 码本设计问题中，不仅仅要设计一个性能较好的多维星座图，同时要保证各个用户能够相对独立地进行发送和接收，才能满足实际系统的应用需求。其次 SCMA 要面临接收机设计。由于不同的层可能会占用相

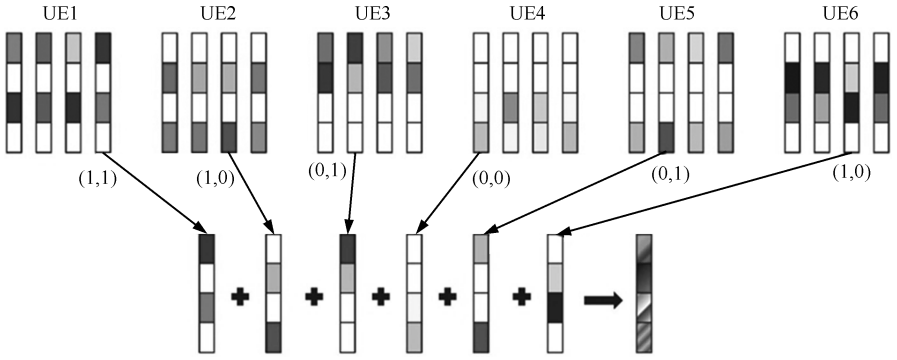


图 12　SCMA 原理示意图

同的时间（或频率）资源，SCMA 中区分用户就成为了技术难点。通常，在接收端使用一个非线性的多用户检测器（multi-user detector，MUD）来区分不同的、非正交的用户，最大似然（maximum likelihood，ML）准则是多用户检测的最优准则。然而，ML 准则的复杂度往往非常之高。考虑到 SCMA 的稀疏特性，可以利用消息传播算法来完成多用户检测。SCMA 的接收问题一直是该领域中热门的研究方向[33]。

频选衰落信道 OFDM-SCMA 重复编码多天线分集系统误码率如图 13 所示。

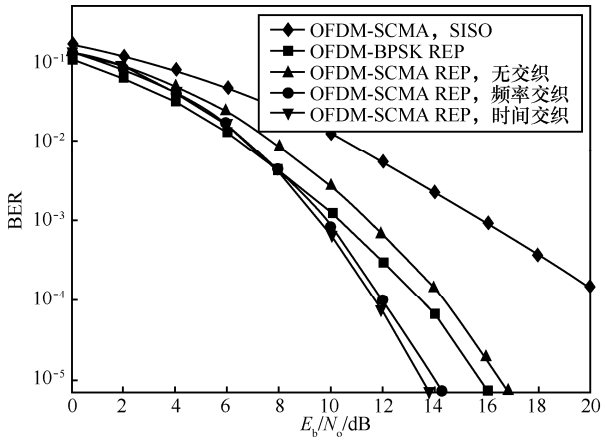


图 13　频选衰落信道 OFDM-SCMA 重复编码多天线分集系统误码率

从图 13 中可以看出，使用重复编码多天线技术的 OFDM-SCM 系统的性能优于单输入单输出 OFDM-SCMA 系统在衰落信道下的性能，说明 MIMO 技术的引入为系统增加了可靠性。比较 SCMA 多天线系统与传统的正交多天线系统可以看出，在低信噪比时，SCMA 方案差于正交方案；同时还能看出，在不使用交织技术时，基于 OFDM-SCMA 的重复编码多天线传输系统的性能比使用 BPSK 调制的 OFDM 传输系统要差。这是因为 SCMA 码本的非正交特性为系统引入了干扰，使其性能本身比正交方案略差。使用频率交织技术后，同一 SCMA 信号的 K 个不同的投影可以映射到近似衰落独立的 OFDM 子载波上，从而产生了频率分集作用。虽然使用频率交织技术的误码率略高于使用时间交织技术的误码率，但这是仿真假设条件所带来的差距。因为在仿真中，假设不同相干时间内的的信道系数是完全独立的，但 OFDM 子载波间的衰落系数是存在一定的相关性的。因此在实际的 OFDM-SCMA 系统中，频率交织和时间交织孰优孰劣要分具体情况讨论。

除了上述的 NOMA 和 SCMA，非正交多址接入技术还有中兴的多用户共享多址（multi-user shared access，MUSA）、大唐的图样分割多址（pattern division multiple access，PDMA）以及高通提出的资源扩频多址（resource spread multiple access，RSMA）等方案。在万物互联的 5G 愿景下，非正交多址接入技术绝对是一场多址技术的改革和创新，会使未来移动通信的无线接入技术达到一个新的高度。

5　结束语

本文首先从 5G 的发展开始介绍了 5G 需要面临的挑战和热点先进技术，然后分别从大规模天线阵列、超密集组网和非正交多址接入技术 3 个方面，对具有 5G 特色技术的研究热点和难点详细进行了阐述。5G 虽然面对了业务需求和用户体验两方面双重的挑战和考验，但各研究机构都积极地为 5G 的铺设进行着努力。距离预计 5G 商用的 2020 年已经不远了，然而 5G 标准的确定之路才刚刚开始，对于 5G 先进技术的研究仍还有大量工作需要完成。在不懈的努力后，未来的 5G 技术定会是一个更开放、更智能、更灵活、更丰富的移动通信技术。

参考文献：

[1] IMT-2020(5G)推进组. 5G 概念白皮书[R]. 2015.
IMT-2020 (5G) Propulsion Group. 5G concept white paper[R]. 2015.

[2] ITU-R. IMT vision, framework and overall objectives of the future development of IMT for 2020 and beyond: ITU-R Document 5/199-E. M 2083-0[S]. 2015.

[3] 张平, 陶运铮, 张治. 5G 若干关键技术评述[J]. 通信学报, 2016, 37(7): 15-29.
ZHANG P, TAO Y Z, ZHANG Z. Survey of several key technologies for 5G[J]. Journal on Communications, 2016, 37(7): 15-29.

[4] AGYAPONG P, IWAMURA M, STAEHLE D, et al. Design considerations for a 5G network architecture[J]. IEEE Communications Magazine, 2014, 52(11): 65-75

[5] 王祖阳, 杨传祥, 张进, 等. 5G 无线网技术特征及部署应对策略分析[J]. 电信科学, 2018, 34(Z1): 9-16.
WANG Z Y, YANG C X, ZHANG J, et al. Analysis of 5G wireless network technology characteristics and deployment strategy[J]. Telecommunications Science, 2018, 34(Z1): 9-16.

[6] 尤肖虎, 潘志文, 高西奇, 等. 5G 移动通信发展趋势与若干关键技术[J].中国科学: 信息科学, 2014, 44(5): 551-563.
YOU X H, PAN Z W, GAO X Q, et al. The 5G mobile communication: the development trends and its emerging key techniques[J]. Scientia Sinica Informationis, 2014, 44(5): 551-563.

[7] ANDREWS J G. What will 5G be?[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(6): 1065-1082.

[8] MARZETTA T. Noncooperative cellular wireless with unlimited numbers of base station antennas[J]. IEEE Transactions on Wireless Communications, 2010,11(9): 3590-3600.

[9] SOHRABI F, YU W. Hybrid digital and analog beamforming design for large-scale antenna arrays[J]. IEEE Journal of Selected Topics in Signal Processing, 2016, 10(3): 501-513.

[10] GAO X, DAI L, CHEN Z, et al. Near-optimal beam selection for beamspace mmWave massive MIMO systems[J]. IEEE Communications Letters, 2016, 20(5): 1054-1057.

[11] SONG N, YANG T, SUN H. Overlapped subarray based hybrid beamforming for millimeter wave multiuser massive MIMO[J]. IEEE Signal Processing Letters, 2017, 24(5): 550-554.

[12] BRADY J, BEHDAD N, SAYEED A M. Beamspace MIMO for millimeter-wave communications: system architecture, modeling, analysis, and measurements[J]. IEEE Transactions on Antennas and Propagation, 2013, 61(7): 3814-3827.

[13] SAYEED A, BRADY J. Beamspace MIMO for high- dimensional multiuser communication at millimeter-wave frequencies[C]//2013 IEEE Global Communications Conference (GLOBECOM), Dec 9-13, 2013, Atlanta, GA, USA. Piscataway: IEEE Press, 2013: 3679-3684.

[14] ZHU X, WANG Z, DAI L, et al. Adaptive hybrid precoding for multiuser massive MIMO[J]. IEEE Communications Letters, 2016, 20(4): 776-779.

[15] SOHRABI F, YU W. Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays[J]. IEEE Journal on Selected Areas in Communications, 2017, 35(7): 1432-1443.

[16] BUZZI S, D'ANDREA C. Energy efficiency and asymptotic performance evaluation of beamforming structures in doubly massive MIMO mmWave systems[J]. IEEE Transactions on Green Communications and Networking, 2018, 2(2): 385-396.

[17] NGUYEN D H N, LE L B, LE-NGOC T, et al. Hybrid MMSE precoding and combining designs for mmWave multiuser systems[J]. IEEE Access, 2017 (5): 19167-19181.

[18] AYACH O E, RAJAGOPAL S, ABU-SURRA S, et al. Spatially sparse precoding in millimeter wave MIMO systems[J]. IEEE Transactions on Wireless Communications, 2014, 13(3): 1499-1513.

[19] LIANG L, XU W, DONG X. Low-complexity hybrid precoding in massive multiuser MIMO systems[J]. IEEE Wireless Communications Letters, 2014, 3(6): 653-656.

[20] SUN W, YUE Q, MENG W, et al. Transmission mechanism and performance analysis of multiuser opportunistic beamforming in Rayleigh and Rician fading channels[J]. IEEE Transactions on Vehicular Technology, (has been accepted).

[21] SADR S, ADVE R S. Partially-distributed resource allocation in small-cell networks[J]. IEEE Transactions on Wireless Communications, 2014, 13(12): 6851-6862.

[22] ANDREWS J G, BACCELLI F, GANTI R K. A tractable approach to coverage and rate in cellular networks[J]. IEEE Transactions on Communications, 2011, 59(11): 3122-3134.

[23] SOKUN H U, BEDEER E, GOHARY R H, et al. Optimization of discrete power and resource block allocation for achieving maximum energy efficiency in OFDMA networks[J]. IEEE Access, 2017 (5): 8648-8658.

[24] HU B, WANG Y, WANG C. A maximum data transmission rate oriented dynamic APs grouping scheme in user-centric UDN[C]//2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Sept 25-26, 2017, Jeju Island, South Korea. Piscataway: IEEE Press, 2017: 56-61.

[25] BOLJEVIC S. Cost of optimal placement of a CHP plant within existing UDN[C]//2017 14th International Conference on the European Energy Market (EEM), June 6-9, 2017, Dresden, Germany. [S.l.:s.n.], 2017: 1-6.

[26] TAO Y, LIU L, LIU S, et al. A survey: several technologies of non-orthogonal transmission for 5G[J]. China Communications, 2015, 121(10): 1-15.

[27] 胡显安. 5G 新型非正交多址技术研究[D]. 北京: 北京交通大学, 2017.
HU X A. Research on 5G new non orthogonal multiple access technology[D]. Beijing: Beijing Jiaotong University, 2017.

[28] HONG S, BRAND J, CHOI J I, et al. Applications of self-interference cancellation in 5G and beyond[J]. IEEE Communications Magazine, 2014, 52(2): 114-121.

[29] 张德坤. 非正交多址系统功率分配及干扰消除算法研究[D]. 哈尔滨: 哈尔滨工业大学, 2015.
ZHANG D K. Algorithm of power allocation and interference cancellation for non-orthogonal multiple access systems[D].

Harbin: Harbin Institute of Technology, 2015.
[30] NIKOPOUR H, BALIGH H. Sparse code multiple access[C]//2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Sept 8-11, 2013, London, UK. Piscataway: IEEE Press, 2013: 332 - 336.
[31] FORNEY G D, WEI L F. Multidimensional constellations. Part. I: introduction, figures of merit, and generalized cross constellations[J]. IEEE Journal on Selected Areas in Communications, 1989, 7(6): 877 - 892.
[32] FORNEY G D. Multidimensional constellations. Part. II: voronio constellations[J]. IEEE Journal on Selected Areas in Communications, 1989, 7(6): 941- 958.
[33] BAYESTEH A, NIKOPOUR H, TAHERZADEH M, et al. Low complexity techniques for SCMA detection[C] //2015 IEEE GLOBECOM Workshops (GC Wkshps), Dec 6-10, 2015, San Diego, CA, USA. Piscataway: IEEE Press, 2015: 1 - 6.
[34] TAHERZADEH M, NIKOPOUR H, BAYESTEH A, et al. SCMA for downlink multiple access of 5G wireless networks[C]//2014 IEEE Global Communications Conference (GlobeCom), Dec 8-12, 2014, Austin, TX, USA. Piscataway: IEEE Press, 2014: 3940 - 3945.
[35] AU K, ZHANG L, NIKOPOUR H, et al. Uplink contention based SCMA for 5G radio access[C]//2014 GlobeCom Workshops (GC Wkshps), Dec 8-12, 2014, Austin, TX, USA. Piscataway: IEEE Press, 2014: 900 - 905.

[作者简介]

林泓池（1993−），男，哈尔滨工业大学硕士生，主要研究方向为超密集组网和非正交多址。

孙文彬（1990−），男，哈尔滨工业大学博士生，主要研究方向为无线通信、多天线技术以及预编码技术。

郭继冲（1992−），男，哈尔滨工业大学博士生，主要研究方向为无线信道建模、毫米波系统、预编码技术。

麻津铭（1994−），男，哈尔滨工业大学硕士生，主要研究方向为超密集组网。

周永康（1994−），男，哈尔滨工业大学硕士生，主要研究方向为稀疏码多址接入的码本设计和技术应用。

于启月（1982−），女，博士，哈尔滨工业大学电子与信息工程学院教授、博士生导师，主要研究方向为宽带无线通信、信息论与编码等。

孟维晓（1968−），男，博士，哈尔滨工业大学电子与信息工程学院教授、博士生导师，主要研究方向为无线移动通信、空天通信网络和卫星定位导航。

专题：5G

基于人工智能的无线传输技术最新研究进展

张静[1]，金石[1]，温朝凯[2]，高飞飞[3]，江涛[4]
（1. 东南大学移动通信国家重点实验室，江苏 南京 210096;
2. 台湾中山大学通讯工程研究所，台湾 高雄 000800; 3. 清华大学自动化系，北京 100084;
4. 华中科技大学武汉光电国家研究中心，湖北 武汉 430074）

摘　要：智能通信被认为是5G之后无线通信发展的主流方向之一，其基本思想是将人工智能引入无线通信系统的各个层面，实现无线通信与人工智能技术的有机融合。目前，该方面研究正在向物理层快速推进，无线传输技术与人工智能的融合还处于初步探索阶段。面向基于人工智能的无线传输关键技术，从信道估计、信号检测、信道状态信息反馈与重建、信道解码、端到端的无线通信系统方面展开了详细介绍，阐述了近年来国际学术界在该方向的最新研究进展，并在此基础上对利用人工智能的无线传输技术发展趋势进行了进一步展望。

关键词：人工智能；无线传输技术；深度学习
中图分类号：TN929.5
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018234

An overview of wireless transmission technology utilizing artificial intelligence

ZHANG Jing[1], JIN Shi[1], WEN Chaokai[2], GAO Feifei[3], JIANG Tao[4]
1. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
2. Institute of Communications Engineering, Sun Yat-Sen University, Kaohsiung 000800, China
3. Department of Automation, Tsinghua University, Beijing 100084, China
4. Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: Intelligent communication is considered to be one of the mainstream directions in the development of wireless communications after 5G. Its basic idea is to introduce artificial intelligence into all aspects of wireless communication systems, realizing the significant integration of wireless communication and artificial intelligence technology. At present, the research in this field is advancing to the physical layer. The combination of wireless transmission and deep learning is also in the preliminary exploration stage. Key technologies of wireless physical layer based on deep learning were introduced in detail from the aspects of channel estimation, signal detection, feedback and reconstruction of channel state information, channel decoding, end-to-end wireless communication systems, the latest research progress of international academic circles in recent years was presented. On this basis, the development trend in the future was preliminarily forecasted.

Key words: artificial intelligence, wireless transmission, deep learning

收稿日期：2018−07−01；修回日期：2018−08−09
基金项目：国家自然科学基金杰出青年科学基金资助项目（No.61625106）；国家自然科学基金资助项目（No.61531011）
Foundation Items: NSFC for Distinguished Young Scholars of China (No.61625106), The National Natural Science Foundation of China (No.61531011)

1　引言

自 2010 年以来，5G 技术备受学术界和工业界的关注，其主要特点为高维度、高容量、更密的网络、更低的时延。相比于已经商用化的 4G 系统，5G 无线传输速率提升 10~100 倍，峰值传输速率达到 10 Gbit/s，端到端时延减至毫秒级，连接设备密度增加 10~100 倍，流量密度提高 1 000 倍，频谱效率提升 5~10 倍，能够在 500 km/h 的速度下保证用户体验。与面向人与人通信的 2G/3G/4G 不同，5G 在设计之初，就考虑了人与人、人与物、物与物的互连。国际电信联盟发布的 5G 八大指标包括：基站峰值速率、用户体验速率、频谱效率、流量空间容量、移动性能、网络能效、连接密度和时延。

迄今为止，5G 主要从 3 个维度实现上述指标，即空口增强、更宽的频谱以及网络密集化。这 3 个维度最具代表性的使能技术分别对应大规模 MIMO（multiple-input multiple-output，MIMO）、毫米波通信以及超密集组网。大规模 MIMO 因具备提升系统容量、频谱效率、用户体验速率、增强全维覆盖和节约能耗等诸多优点，被认为是 5G 最具潜力的核心技术。然而，大规模 MIMO 的发展和应用也面临诸多问题，如对于不具有上下行互易性的频分双工（frequency division duplex，FDD）系统，如何有效地实现基站侧的信道状态信息获取。毫米波是指波长在毫米数量级的电磁波，其频率大约在 30~300 GHz。现有的无线通信系统所用到频段大多集中在 300 MHz~3 GHz，对毫米波段的利用率较低。毫米波技术通过增加频谱带宽，有效提高网络传输速率，但会受传播路径损耗、建筑物穿透损耗和雨衰等因素的影响，在实际应用中面临着巨大挑战[1]。另外，毫米波通信可与大规模 MIMO 有机融合，通过大规模 MIMO 波束成形带来的增益弥补毫米波穿透力差的劣势。超密集组网（ultra dense network，UDN）通过更加“密集化”的无线网络部署，将站间距离缩短为几十米甚至十几米，使得站点密度大大增加，从而提高频谱复用率、单位面积的网络容量以及用户体验速率。综合来看，大规模 MIMO 利用超高天线维度充分挖掘利用空间资源，毫米波通信利用超大带宽提升网络吞吐量，超密集组网利用超密基站提高频谱利用率，由此产生了海量的无线大数据，为未来无线通信系统利用人工智能手段提供了数据来源。

另一方面，近年来人工智能特别是深度学习在计算机视觉、自然语言处理、语音识别等领域获得了巨大成功[2]，无线通信领域的研究者们期望将其应用于系统的各个层面，进而产生智能通信系统，实现真正意义上的万物互联，满足人们对数据传输速率日新月异的需求。因此，智能通信被认为是 5G 之后无线通信发展的主流方向之一，其基本思想是将人工智能引入无线通信系统的各个层面，实现无线通信与人工智能技术的有机融合，大幅度提升无线通信系统的效能。学术界和工业界正在上述领域开展研究工作，前期的研究成果集中于应用层和网络层，主要思想是将人工智能特别是深度学习的思想引入无线资源管理和分配等领域。目前，该方向的研究正在向 MAC 层和物理层推进，特别在物理层已经出现无线传输与深度学习等结合的趋势，然而，各项研究目前尚处于初步探索阶段。

尽管无线大数据为人工智能应用于物理层提供可能[3]，智能通信系统的发展仍处于探索阶段，机遇与挑战并存。追溯历史，无线通信系统从 1G 演进至 5G 并获得巨大成功，其根源在于基于香农信息论的无线传输理论体系架构的建立与完善。一个典型的无线通信系统由发射机、无线信道和接收机构成，如图 1 所示。发射机主要包括信源、信源编码、信道编码、调制和射频发送等模块；接收机包括射频接收、信道估计与信号检测、解调、信道解码、信源解码以及信宿等模块。智能

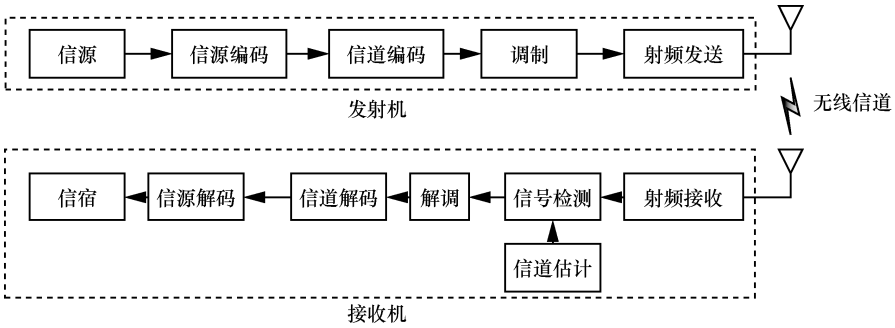


图 1　典型的无线通信系统

通信的无线传输研究旨在打破原有的通信模式，获得无线传输性能的大幅提升。目前这方面的研究面临诸多挑战，国内外研究者进行了初步探索。本文主要介绍深度学习应用于无线传输技术的最新研究进展，主要包括信道估计、信号检测、信道状态信息（channel state information，CSI）的反馈与重建、信道解码以及端到端的通信系统。

2　深度学习简介

1996 年 Langley 将机器学习定义为人工智能的一个分支，旨在依赖经验知识提高系统性能。经过 20 世纪以来的长期研究，研究者提出了逻辑回归、决策树、支持向量机和神经网络等各种算法。2006 年，Hinton 等人[4]在《Science》上发表论文，其主要观点有：多隐层的人工神经网络具有优异的特征学习能力；可通过“逐层预训练”来有效克服深层神经网络在训练上的困难，从此引出深度学习的研究。而后，深度学习在语音识别领域和图像识别领域取得巨大成就。深度学习作为一种新兴的神经网络算法，具有多种结构，包括深度神经网络（deep neural network，DNN）、卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）和生成对抗神经网络（generative adversarial network，GAN）等。下面详细介绍 4 类深度学习网络的基本结构。

2.1　深度神经网络

DNN 也被称为多层感知机。DNN 基本结构如图 2 所示，由输入层、多个隐藏层和输出层构成。每个隐藏层包含多个神经元，每个神经元连接到相邻的层，同层神经元互不连接。单个神经元将各个输入与相应权重相乘，然后加偏置参数，最后通过非线性激活函数。激活函数类型见表 1。反向传播算法是一种有效的 DNN 优化方法，隐藏层和神经元数量的增加使得训练过程变得困难，会遇到如梯度消失、收敛缓慢以及收敛到局部最小值等问题。为了解决消失梯度问题，引入了新的激活函数来代替经典的 Sigmoid 函数。为了提高收敛速度和降低计算复杂度，经典梯度下降法（gradient descent，GD）被调整为随机梯度下降法（stochastic gradient descent，SGD），它随机选择一个样本来计算每次的损失和梯度。随机特性在训练过程中会引起强烈的波动，因此，在经典的 GD 和 SGD 之间采用小批量随机梯度下降法（small-batch SGD）进行训练。然而，这些算法仍然会出现收敛于局部最优解。为了解决这一问题并进一步提高训练速度，几种自适应学习速率算法应运而生，如 Adagrad、RMSProp、Momentum、Adam 等[4]。如果训练后的网络在训练数据上表现良好，在测试过程中表现不佳，则出现过度拟合现象。在这种情况下，为了在训练和测试数据上取得良好的结果，提出了正则化（regularization）和丢弃（dropout）等方案。

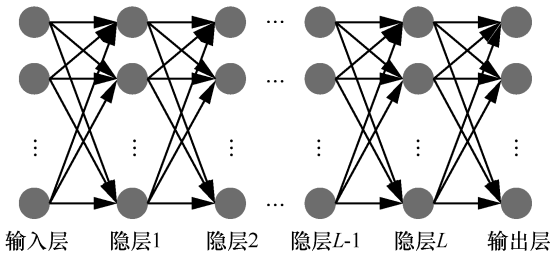


图 2　DNN 基本结构

表 1　激活函数类型

名称	激活函数
Sigmoid	$\frac{1}{1+e^{-x}}$
Tanh	$\tanh(x)$
ReLu	$\max(0,x)$

2.2　卷积神经网络

CNN 的基本结构包括输入层、多个卷积层、多个池化层、全连接层及输出层，如图 3 所示。卷积层和池化层采用交替设置，即一个卷积层连接一个池化层，池化层后再连接一个卷积层，依此类推。由于卷积层中卷积核的每个神经元与其输入进行局部连接，并通过对应的连接权值与局部输入进行加权求和再加入偏置值，得到该神经元输出值，该过程等同于卷积过程，因此被称为卷积神经网络。

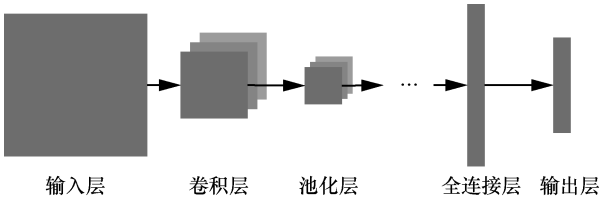


图 3　CNN 基本结构

2.3　循环神经网络

RNN 是一种对序列数据建模的神经网络，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对过去时刻的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再是无连接的而是有连接的，并且隐藏层的输入不仅包括输入层还包括上一时刻隐藏层的输出。图 4 是一个 RNN 模型的示例。循环神经网络旨在为神经网络提供记忆，因为输出不仅依赖于当前输入，而且还依赖于过去时刻可用的信息或将来时刻可用的信息。图 4 所示的时延步长（time step）为 3。常用的 RNN 包括 Elman 网络、Jordan 网络、双向 RNN 和长短时记忆（long short term memory，LSTM）。

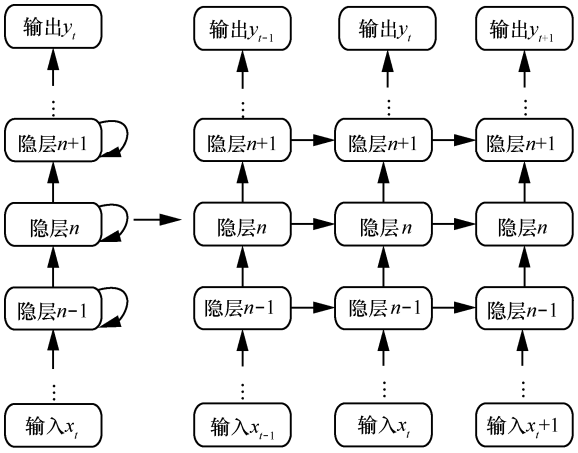


图 4　RNN 基本结构

2.4　生成对抗神经网络

GAN 是一种新型的分布学习生成方法，目的是学习一种能够在真实分布的数据集上生成伪样本的模型。GAN 的结构如图 5 所示，包含一个生成器 G 和一个鉴别器 D。生成器和鉴别器均由 DNN 实现。鉴别器用于区分生成器生成的伪样本和实际数据集的真样本，生成器任务是生成样本数据使得鉴别器区分不出真样本和伪样本。在训练过程中，生成器将输入噪声 z 与样本的先验分布 $p_z(z)$映射到一个样本。然后采集来自真实数据的样本和来自于生成器 G 的样本，以训练鉴别器 D，以最大化区分这两类的能力。如果鉴别器 D 成功地对真样本和假样本进行分类，那么它的成功可以反馈给生成器 G，从而促使生成器 G 学会生成与真样本更相似的样本。训练过程在达到平衡时结束，此时鉴别器 D 只能随机猜测真样本和产生的伪样本。

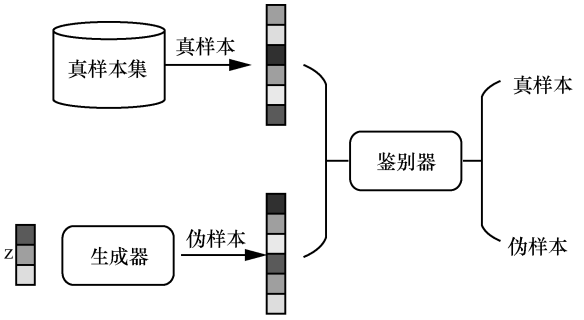


图 5　GAN 基本结构

3　深度学习在无线传输技术中的应用

本节面向基于深度学习的无线物理层关键技术，从信道估计、信号检测、CSI 反馈与重建、信道解码以及端到端无线通信系统 5 个方面展开详细介绍，展示了近年来国际学术界在该方面的最新研究进展。

3.1　信道估计

在大规模 MIMO 波束毫米波场景下，信道估计极具挑战性，尤其是在天线阵列密集、接收机配备的射频链路受限的场景。参考文献[5]提出 LDAMP 网络来解决这一信道估计问题。该网络将信道矩阵视作二维图像作为输入，并将降噪的卷积神经网络融合到迭代信号重建算法中进行信道估计。LDAMP 基于 D-AMP 算法[6]，由 L 层完全相同的结构串联而成。每层由降噪器、散度估量器和连接的系数组成。降噪器由具有 20 个卷积层的 DnCNN 实现，在 LDAMP 网络起到决定性作用。在未知噪声强度情况下，DnCNN 降噪器能够解决高斯降噪问题，比其他降噪技术准确度更高、计算速度更快。它不是直接从含有噪声的信道图像中学习信道图像，而是先学习残余噪声，然后通过相减操作获得信道估计的图像。残差学习不仅降低了训练时间，也增强了信道估计的准确性。仿真结果表明 LDAMP 网络的性能优越于当前最具潜力的其他信道估计方法。

从 MMSE 算法的基本结构出发，参考文献[7]提出了一种基于深度学习的信道估计器，其中估计的信道向量为条件高斯随机变量，协方差矩阵具有随机性。如果协方差矩阵具有特普利兹特性和移不变的结构特性，则 MMSE 信道估计器的复杂度将降低很多。在信道的协方差矩阵不具备上述特性时，信道估计的复杂度将会变得很大。为了降低信道估计的复杂度，参考文献[7]仍假设采用 MMSE 的结构模型，并利用 CNN 对误差进行补偿。仿真结果表明，提出的信道估计器在降低复杂度的同时，也保证了信道估计的准确性。

参考文献[5]和参考文献[7]不仅考虑到实际问题中的模型特点，而且以已有算法为基础，使整个深度学习网络的学习参数较少，而且准确性高、复杂性低，更具竞争力。

3.2　信号检测

参考文献[8]利用 DNN 实现 OFDM（orthogonal frequency division multiplexing，OFDM）系统中的信号检测问题。传统的 OFDM 系统信道估计和信号检测是两个独立的功能模块，即先进行信道估计获得确切的 CSI，然后利用估计的 CSI 对发送信号进行恢复。如图 1 所示，信号原始恢复过程还涉及解调等模块。与传统无线通信不同，参考文献[8]将信道估计和信号检测视为一个整体，直接用 DNN 实现由接收信号到原始信号的映射。DNN 的输入为 256，隐层结构为 500-250-120，输出为 16。文中的 OFDM 系统采用 64 子载波，调制方式为正交相移（quadrature phase shift keying，QPSK）编码，因此输入信号为 128 byte，因此需要 8 个相同结构的 DNN 进行训练。所提出的 DNN 在训练大量数据以后，能与传统检测算法最小均方误差（minimum mean square error，MMSE）性能相比拟。在无循环前缀或降低峰均信噪比的 OFDM 非线性系统中，此 DNN 获得的性能比传统的 MMSE 提升很多。然而，这种提升并非表明这种设计的合理性，误比特率也呈现一定的饱和性。饱和性是指随着信噪比的增大，信号检测的误比特率不再下降或下降不明显。在实际系统中，

非线性情况还没有得到很好的解决。另外需要提出的是一个子 DNN 需要训练 20 000 次，一次训练涉及 50 000 个数据。8 个这样的网络其训练时间和复杂度可想而知。

参考文献[9]研究的是 MIMO 系统的信号重建问题，提出了信号检测算法 DetNet。DetNet 在最大似然法基础上加入梯度下降算法，从而生成一个深度学习网络。为了测试 DetNet 的顽健性，考虑了两种 CSI 已知的情景，即时不变信道和随机变量已知的时变信道。仿真结果表明，DetNet 性能优于传统的信号检测算法 AMP（approximate message passing），而且与 SDR（semidefinite relaxation）算法性能相当，具有极高的准确性和极少的时间开销（该算法速度是传统算法的 30 倍）。

参考文献[10]与参考文献[9]研究的问题相同，而且提出的解决方法均依赖已有的信号检测算法。参考文献[10]以 OAMP（orthogonal approximate message passing，OAMP）迭代算法为基础，结合深度学习网络提出了 OAMP-Net，目的是在原有算法基础之上，加入可调节的训练参数，进一步提升已有算法的信号检测性能。OAMP 算法在压缩感知领域被提出来解决稀疏线性求逆问题，后被用于 MIMO 的信号检测问题，与以往算法相比，OAMP 算法复杂度降低很多。但是，信号检测性能有所下降。OAMP 算法是一种迭代算法，增加了算法的复杂度。为了进一步降低算法的复杂度，OAMP-Net 包含了 T 个串联层，相当于算法的迭代过程。每个串联层不仅实现了 OAMP 算法的全过程，而且加入了一些可训练的参数使得 OAMP 算法更具弹性，在参数改变时，不仅能适应更多的信道场景，而且可以实现与其他算法模型的转换。仿真结果表明，OAMP-Net 的性能不仅高于 OAMP 算法，而且优越于更加复杂的 LMMSE-TISTA 算法，其算法复杂度更低且能适应于时变的信道。

不难看出，参考文献[8]提出的深度学习网络需要采集大量训练数据，前期训练工作量巨大，参考文献[9-10]克服了这一困难，利用更简单的训练网络实现了更好的信号检测性能。

3.3　CSI 反馈与重建

在频分复用网络中，MIMO 系统中基站需要获得下行链路的 CSI 反馈来执行预编码以及实现性能增益。然而 MIMO 系统中的超多天线造成过量的反馈负载，因此传统的 CSI 反馈负载降低方法不再适用于此场景。参考文献[11]提出基于 CNN 的 CSI 感知与恢复机制 CsiNet。CsiNet 的感知部分也称为编码器，将原始 CSI 矩阵利用 CNN 转化为码本；CsiNet 的恢复部分也称为译码器，将接收到的码本利用全连接网络和 CNN 恢复成原始的 CSI 信号。编码器网络包括 32×32 输入层、2 个 3×3 卷积核、1×N 重建层（reshape）和一个线性的 1×M 全连接层。解码器网络包括 1×M 输入层、1×N 全连接层、32×32 重建层和两个 Refine 网络。Refine 网络包括 4 层 3×3 卷积层进行特征提取。参考文献[12]在压缩 CSI 反馈的空间复用 MIMO 系统中只是利用 DNN 将原始 CSI 矩阵压缩为低维度的 CSI 信号，没有涉及 CSI 反馈信号的进一步恢复。

参考文献[13]在参考文献[11]基础上提出一种实时的基于 LSTM 的 CSI 反馈架构 CsiNet-LSTM，该网络利用 CNN 和 RNN 分别提取 CSI 的空间特征和帧内相关性特征，从而进一步提升反馈 CSI 的正确性。CsiNet-LSTM 的时延步长为 T，第一步时延步长的信道矩阵采用高压缩率编码器，其他 T-1 个时延步长采用低压缩率编码器。T-1 个低压缩率编码器的输出码字分别与高压缩率编码器的输出码字串联在一起，然后输入相应的译码器中。最后的 CSI 重建由 T 个时延步长具有 3 层 2×32×32 单元的 LSTM 执行。需要指出的是，此网络编码器和译码器部分与参考文献[11]的 CsiNet 结构完全相同。利用时变 MIMO

信道时间相关性和结构特点，CsiNet-LSTM 能实现压缩率、CSI 重建质量以及复杂度之间的折中。相比于 CsiNet，该网络以时间效率换取了 CSI 的重建质量。

参考文献[11,13]提出的 CSI 反馈与重建算法均依赖大量数据进行离线训练，网络复杂度较高且泛化性能需要深入研究。

3.4 信道解码

参考文献[14]提出了一种基于 DNN 的信道解码方法。该文献得出了深度学习应用于信道解码的两个结论，一是如极化码等结构码比随机码更容易学习；二是针对结构码，深度学习网络能够解码没有训练过的码字。在接收端，k 位信息比特被编码为长度为 N 的码字，然后进行调制，发射机通过噪声信道将其送至接收端。接收端的信道解码器的任务是将接收的具有噪声干扰的码字恢复成相应的信息比特。信道解码器由输入层、3 层隐层、输出层构成。输入层为具有噪声干扰的码字，输出层为信息比特。3 层隐层神经元结构为 128-64-32。信道解码深度学习用于信道解码无疑会受到维度爆炸的限制，对于码长为 100，码率为 0.5 的编码来说，则存在 2^{50} 种不同的码字。因此，该网络只适应于码字较短的信道编码技术。仿真结果表明，对于结构码来说，训练 2^{19} 次则接近最大后验概率（maximum a posteriori，MAP）解码器的性能。而对于随机码来说，训练 2^{19} 次性能远远不及 MAP 解码器。另外，参考文献[14]分别对不同的隐层结构 128-64-32、256-128-64、512-256-128 和 1 024-512-128 进行了比较，对于此解码网络来说，隐层结构越复杂，训练的次数越多，该解码器的性能越优越。

参考文献[14]将解码部分视作一个黑盒子，直接实现了从接收码字到信息比特的转换，该方式的性能虽然能与传统方法相比拟，但是训练次数呈指数上升，深度学习网络结构也足够复杂，当码长发生改变时，该网络需要被重新调整输入输出，并重新训练，工作量之大可想而知。另外，该方法不适用于随机码，也不适合码长较长的码字，具有很大的局限性。参考文献[15]与参考文献[14]不同，在传统极化码迭代解码算法基础上，提出一种分离子块的深度学习极化码解码网络。该网络主要包含两个步骤：一是将原编解码分割成 M 个子块，然后分别对各个子块进行编码/解码，子块解码过程采用 DNN，性能接近 MAP 解码器的性能，子块的引入克服了码长过长造成的解码复杂度问题；二是利用置信传播（belief propagation，BP）解码算法连接各个子块，BP 算法与子块 DNN 连接实现并行处理。参考文献[15]的解码算法是一个高度并行的解码算法，且不属于迭代算法。该算法[15]与传统算法相比，算法时延大大降低，且性能相当；与参考文献[14]的解码算法相比，该算法在训练次数和网络结构上的复杂度均大大降低。

参考文献[16]与参考文献[15]均利用 BP 算法与深度学习网络结合进行信道解码。参考文献[16]提出一种迭代的信道解码算法 BP-CNN，该算法将 CNN 与标准 BP 解码器串联，在噪声环境中估计信息比特。在接收端，接收到的信号首先由 BP 解码器进行处理以获得原始的解码结果，然后用接收信号与估计的传输符号相减获得信道噪声估计。由于存在解码误差，信道噪声估计具有较大误差。最后将信道噪声估计输入 CNN 以移除 BP 解码器的估计误差。标准 BP 接收机用于估计传输信号，CNN 用于降低 BP 检测器的估计误差，并且获得更加确切的信道噪声估计。BP 算法与 CNN 之间的迭代会逐渐提高检测 SNR，因此获得更好的译码性能。BP-CNN 不仅解码性能优于标准 BP 算法，而且具有更低的复杂度。这是因为 CNN 的高效性，CNN 大部分操作为线性运算，少部分为非线性运算。仿真结果表明，噪声相关性越强，BP-CNN 的性能优越性越高，当噪声非相关时，BP-CNN 的性能稍逊于 BP 算法。

对于高密度奇偶校验码（high density parity check，HDPC）来说，BP 算法的性能相对较差。Tanner 图为校验码的校验矩阵，Nachmani 等人[17]提出了 BP-DNN 算法，该算法为 Tanner 图的边缘节点分配权重系数，并利用 DNN 进行训练获得这些系数，从而提升 BP 算法应用于 HDPC 的性能。BP-DNN 的迭代次数约为 BP 算法的 1/10，且性能优于传统的 BP 算法。Nachmani 等人[18]又提出 BP-RNN 算法，将 RNN 网络与 BP 算法结合，进一步提升 BP 算法的性能。另外，参考文献[18]将 BP-RNN 中的 BP 算法替换为改进的随机冗余迭代算法（modified random redundant iterative algorithm，mRRD）算法，获得的解码性能优于 mRRD 算法。

上述基于深度学习的信道解码方法中，参考文献[14]将无线通信系统中解码模块看作黑盒子，其他文献均与 BP 算法结合进一步寻求性能的提升以及复杂度的降低。

3.5　端到端无线通信系统

O'Shea 等人[12]提出一种 MIMO 系统中基于深度学习自编码器的物理层策略。在特定的信道环境下，此策略利用自编码器对估计、反馈、编码和解码过程进行全局优化，以达到最大化吞吐量和最小化误比特率的目的。参考文献[12]利用自编码器实现了 3 个无线通信系统，分别是：无 CSI 反馈的空间分集 MIMO 系统，完美 CSI 反馈的空间复用 MIMO 系统以及压缩 CSI 反馈的空间复用 MIMO 系统。在特定信道环境下，此物理层策略获得了较大的性能提升。

参考文献[19]提出了一个点对点无线通信系统模型，诠释了物理层的处理模块由 DNN 替代的可行性。传统无线通信系统的设计需要考虑硬件实现时各种不确定因素的影响，并进行时延、相位等方面的补偿。其[19]提出的基于 DNN 的端到端无线通信系统也考虑了这些因素，在系统实现时进行两个阶段的训练。第一阶段为随机信道下的发送、信道与接收 DNN 的训练。第二个阶段在第一阶段网络训练参数的基础上，在真实信道下再一次进行训练，对训练参数进行微调，使得整个系统的性能进一步提升。信道模块中，时延和相位补偿均被考虑到 DNN 的训练中。接收模块中，接收信号特征提取和相位补偿由 DNN 替代，两个 DNN 训练结果串联起来输入接收 DNN 中。这种基于 DNN 的无线通信系统充分考虑了真实信道下的时变性，系统性能与传统无线通信系统性能具有可比性。

参考文献[20]利用 DNN 实现端到端无线通信系统，其中信号相关的功能模块均利用 DNN 实现，如编码、解码、调制以及均衡。无线通信系统中瞬时 CSI 很难准确获取，而且随着时间和位置的改变不断变化。整个端到端的无线通信系统在反向传播计算梯度时由于信道未知无法进行。参考文献[20]提出一种信道未知情况下的端到端系统，它不依赖任何信道的先验知识。系统采用 GAN 代表无线信道影响，发送端的编码信号作为条件信息。为了克服信道的时变性，导频数据的接收信号也作为条件信息的一部分。此无线通信系统发送机和接收机各由一个 DNN 代替，GAN 作为发送端与接收端的桥梁，使得反向传播顺利进行。发送 DNN、接收 DNN、信道生成 GAN 相互迭代进行训练，最终得到全局最优解。仿真结果表明，利用 GAN 进行信道估计的方法与传统信道估计性能相当，端到端无线通信系统的性能接近传统的基于通信知识的信道模型系统。此方法打破了传统模型化的无线通信模式，为无线通信系统设计开辟新道路。

基于端到端的无线通信系统也被称为自编码器，用编码、信道、解码过程代替原先的无线通信系统结构，编码、信道、解码部分均用深度学习网络实现，是一种全新的无线通信系统实现思路。然而，多个 DNN 需要依赖大量的数据，数据的采集过程任务量巨大，如果环境或硬件系统发生改变，信息的采集过程需要重新进行，目前这种方式实现无线通信系统的可操作性较低。

4 总结与展望

5G 技术呈现高维度、高容量、高密集的特点，在无线传输中产生海量数据，物理层中的大数据成为一个兴趣点，期望利用人工智能提升物理层的传输性能。近年来，研究者已经对此做了初步探索，主要呈现出两种类型的深度学习网络，一种基于数据驱动，另一种基于数据模型双驱动。基于数据驱动的深度学习网络[8,11,14,19-20]将无线通信系统的多个功能块看作一个未知的黑盒子，利用深度学习网络取而代之，然后依赖大量训练数据完成输入到输出的训练。例如，参考文献[8]将 OFDM 系统中整个接收模块作为一个黑盒子，射频接收机接收到信号之后，然后进行移除循环前缀操作，最后利用 DNN 直接完成从射频接收机到信宿的过程。基于端到端的无线通信系统将整个通信系统由深度学习网络全面替代，期望全局优化无线通信系统，获得更好的性能[12,20]。基于数据模型双驱动的深度学习网络[5,7,9-10,15,18]在无线通信系统原有技术的基础上，不改变无线通信系统的模型结构，利用深度学习网络代替某个模块或者训练相关参数以提升某个模块的性能。例如，参考文献[6]在无线通信系统 MIMO 信号检测模块 OAMP 接收机基础上，利用深度学习网络引入可训练的参数，进一步提升信号检测模块的性能。基于数据驱动的深度学习网络主要依赖海量数据，而基于数据模型双驱动的深度学习网络主要依赖通信模型或者算法模型。

基于数据驱动的深度学习网络通过大量实例学习，吸收了被人类分析员分别标记的大量数据，以生成期望的输出。然而，训练深度学习网络需要大量的标记数据，积累和标记大量信息的过程不但费时而且成本高昂。除了积累标记数据的挑战之外，大多数基于数据驱动的深度学习模型泛化性和自适应性较弱，即使网络部分结构发生微小变化，也会导致训练模型准确性大大降低。例如，如果参考文献[8]发送端的调制方式更换为 16QAM（quadrature amplitude modulation）或 64QAM，网络需要重新训练。因此，调整或修改模型所耗费的代价相当于重新创建模型。为了减少训练和调整深度学习模型的成本和时间，基于模型的深度学习网络具有更好的泛化性和自适应性。蜂窝移动通信从 1G 演进到 5G，无线通信系统性能的提升离不开功能模块的建模，基于数据驱动的深度学习网络摒弃这些已有的无线通信知识，需要海量数据进行训练与学习，而获得的性能达不到已有无线通信系统模型的性能。而基于数据模型双驱动的深度学习网络以物理层已有模型为基础，可以显著减少训练或升级所需的信息量。由于已有的模型具有环境自适应性和泛化性，因此数据模型双驱动深度学习网络也具有这些特性，并且能在原模型基础上进一步提升系统的性能。数据驱动与数据模型双驱动的深度学习网络比较见表 2。由本文第 3 节的分析可知，数据模型双驱动深度学习网络在信道估计、信号检测、信道解码的应用上取得的良好性能，具有广阔的发展前景。

表 2 数据驱动与数据模型双驱动的深度学习网络比较

类型	数据依赖性	模型依赖性	准确性	复杂度
数据驱动	高	低	相对较低	高
数据模型双驱动	低	高	高	低

5 结束语

本文首先介绍了当前应用较广的深度学习网络的几种类型，包括 DNN、CNN、RNN 和 GAN。然后详细阐述了深度学习应用于无线传输技术的最新研究成果，包括信道估计、信号检测、CSI 反馈与重建、信道解码以及端到端无线通信系统。智能无线通信作为后 5G 发展的主流技术之一，物理层寻求技术上的突破关键在于利用大数据降低系统实现复杂度、提升系统性能。从最新研究进

展中可以看出，数据模型双驱动的深度学习网络不仅能满足这些需求，而且对数据的依赖性大大减小，成为最具潜力的发展方向之一。

参考文献：

[1] 倪善金, 赵军辉. 5G 无线通信网络物理层关键技术[J]. 电信科学, 2015, 31(12): 40-45.
NI S J, ZHAO J H. Key technologies in physical layer of 5G wireless communications network[J]. Telecommunications Science, 2015, 31(12): 40-45.

[2] MAO Q, HU F, HAO Q. Deep learning for intelligent wireless networks: a comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2018(99):1.

[3] O'SHEA T J, HOYDIS J. An introduction to deep learning for the physical layer[J]. arXiv: 1702.008320, 2017.

[4] WANG T Q, WEN C K, WANG H, et al. Deep learning for wireless physical layer: opportunities and challenges[J]. China Communications, 2017, 14(11): 92-111.

[5] HE H, WEN C K, JIN S, et al. Deep learning-based channel estimation for beamspace mmWave massive MIMO systems[J]. IEEE Wireless Communications Letters, 2018(99): 1.

[6] METZLER C A, MOUSAVI A, BARANIUK R G. Learned D-AMP: principled neural network based compressive image recovery[J]. arXiv: 1704.06625, 2017.

[7] NEUMANN D, WIESE T, UTSCHICK W. Learning the MMSE channel estimator[J]. IEEE Transactions on Signal Processing, 2018, 66(11): 2905-2917.

[8] YE H, LI G Y, JUANG B H F. Power of deep learning for channel estimation and signal detection in OFDM systems[J]. IEEE Wireless Communications Letters, 2018, 7(1): 114-117.

[9] SAMUEL N, DISKIN T, WIESEL A. Deep MIMO detection[C]//IEEE International Workshop on Signal Processing Advances in Wireless Communications, Jul 3-6, 2017, Sapporo, Japan. Piscataway: IEEE Press, 2017.

[10] HE H, WEN C K, JIN S, et al. A model-driven deep learning network for MIMO detection[C]//Submitted to the 6th IEEE Global Conference on Signal and Information Processing, Nov 26-29, 2018, Anaheim, USA. Piscataway: IEEE Press, 2018.

[11] WEN C K, SHIH W T, JIN S. Deep learning for massive MIMO CSI feedback[J]. IEEE Wireless Communications Letters, 2018(99).

[12] O'SHEA T J, ERPEK T, CLANCY T C. Deep learning based MIMO communications[J]. arXiv:1707.07980, 2017.

[13] WANG T Q, WEN C K, JIN S, et al. Deep learning-based CSI feedback approach for time-varying massive MIMO channels[J]. arXiv:1807.11673, 2018.

[14] CAMMERER S, HOYDIS J, BRINK S T. On deep learning-based channel decoding[C]//51st Annual Conference on Information Sciences and Systems, March 22-24, 2017, Baltimore, MD, USA. [S.l.:s.n.], 2017.

[15] CAMMERER S, HOYDIS J, BRINK S T. Scaling deep learning-based decoding of polar codes via partitioning[J]. arXiv: 1702.06901, 2017.

[16] LIANG F, SHEN C, WU F. An iterative BP-CNN architecture for channel decoding[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(1): 144-159.

[17] NACHMANI E, BEERY Y, BURSHTEIN D. Learning to decode linear codes using deep learning[C]//54th Annual Allerton Conference on Communication, Control, and Computing, Sept 27-31, 2016, Monticello, Illinois, USA. [S.l.:s.n.], 2016.

[18] NACHMANI E, MARCIANO E, LUGOSCH L, et al. Deep learning methods for improved decoding of linear codes[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(1): 119-131.

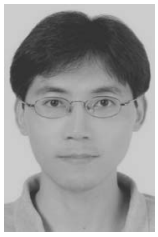
[19] DÖRNER S, CAMMERER S, HOYDIS J, et al. Deep learning based communication over the air[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(1): 132-143.

[20] YE H, LI G Y, JUANG B H F, et al. Channel agnostic end-to-end learning based communication systems with conditional GAN[J]. arXiv: 1807.00447, 2018.

[作者简介]

张静（1993−），女，东南大学移动通信国家重点实验室博士生，主要研究方向为 5G 移动通信物理层关键技术、机器学习等。

金石（1974−），男，东南大学移动通信国家重点实验室教授、博士生导师，主要研究方向为移动通信理论与关键技术、物联网理论与关键技术以及人工智能在无线通信中的应用等。



温朝凯（1976−），男，台湾中山大学通讯工程研究所教授，主要研究方向为无线通信、最优化理论和机器学习等。

高飞飞（1980−），男，清华大学自动化系信息处理研究所副教授，主要研究方向为通信信号处理、大规模多天线技术以及智能通信。

江涛（1970−），男，华中科技大学武汉光电国家研究中心教授、博士生导师，主要研究方向为 5G 移动通信理论与关键技术、天地一体化信息网络、深海目标探测等。

专题：5G

面向商用的 5G 网络关键问题研究及验证

刘玮，董江波，任冶冰
（中国移动通信集团设计院有限公司，北京 100080）

摘　要：3GPP 第一版 5G 标准已于 2018 年 6 月冻结，相关端到端产品也已稳步发展，面向商用的 5G 网络规模试验工作已经启动，以支撑明确技术路线及引入策略。首先介绍了全球 5G 产业发展情况，然后从新技术引入策略、建网方案和业务需求 3 个角度梳理了面向商用的 5G 关键问题，并对 NSA/SA 架构选择、MEC 部署、频率、覆盖与容量、垂直行业业务需求等典型问题进行分析；最后介绍了此次 5G 规模试验的基本情况，包括原则与目标、总体规划与试验内容。

关键词：NSA；SA；传播损耗；规划仿真
中图分类号：TN929.5
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018232

Research and verification of key issues in 5G network

LIU Wei, DONG Jiangbo, REN Yebing
China Mobile Communications Group Design Institute Co., Ltd., Beijing 100080, China

Abstract: The first edition of 5G technical specification has been frozen on June 2018, and the end to end products have also been steadily developing. The 5G network scale test has been launched, to clear strategy for introducing technology and new technologies.The development of 5G industry in the world was introduced. And then the key issues of 5G were combed from 3 dimensions. Finally, the basic situation of the 5G scale test was introduced, including the principles and objectives, the overall plan and the test content.

Key words: NSA, SA, propagation loss, network planning

1　引言

5G 三大主要应用场景分别是：连续广覆盖及高容量场景（eMBB）、低时延高可靠场景（uRLLC）以及低功耗大连接场景（mMTC）。面向 eMBB 场景的 5G 技术框架通过 3GPP R15 版本制定，该版本已于 2018 年 6 月冻结，而面向 uRLLC 和 mMTC 场景的技术方案将在 3GPP R16 版本中制定，除此以外，3GPP R16 版本还将制定一些增强技术方案以持续提升 eMBB 场景的竞争力。

目前，5G 系统侧主设备厂商主要有华为、中兴、诺基亚、大唐和爱立信，终端芯片厂商主要有高通、海思、三星、英特尔。由于 5G 非独立组网架构（non-standalone，NSA）标准冻结较独立

收稿日期：2018−07−22；修回日期：2018−08−10

组网架构（standalone，SA）早半年，因此，主设备厂商和终端芯片厂商对于 NSA 和 SA 两种组网架构的产品研发计划也不同步，NSA 产品研发计划较 SA 产品研发计划早 3~6 个月。当前大部分厂商面向 NSA 组网架构的基站侧设备已于 2018 年第二季度推出，面向 SA 组网架构的基站侧设备也将于 2018 年第三季度推出，面向 NSA 和 SA 组网架构的核心网设备将于 2018 年第三季度和第四季度推出。各厂商终端芯片厂商产品研发技术差异较大，预计在 2018 年第四季度推出面向 NSA 组网架构的终端芯片，在 2019 年第一季度推出面向 SA 组网架构的终端芯片。因此，2018 年第四季度将基本具备面向 NSA 组网架构的 5G 芯片级端到端测试条件，而面向 SA 组网架构的 5G 芯片级端到端测试条件要稍晚。

2　5G 网络商用的关键问题

5G 网络商用关键问题可分为端到端重大方案、无线网、核心网、信令网、传输/承载网、终端、计费、网管编排、能力开放以及安全等 17 大类 82 项，如 5G 语音方案、4G 与 5G 互通策略、4G 演进与 5G 关系、5G 空口安全等。本文主要论述新技术引入策略、建网方案、业务需求这 3 方面涉及的关键问题。

2.1　新技术引入策略

为了满足 eMBB、uRLLC 和 mMTC 三大场景业务需求，与 4G 相比，5G 端到端发生了新的技术变革，如图 1 所示。业务需求不同、产业进展不同、技术灵活多样，5G 网络商用面临的第一大问题便是在 5G 不同的发展阶段，各种

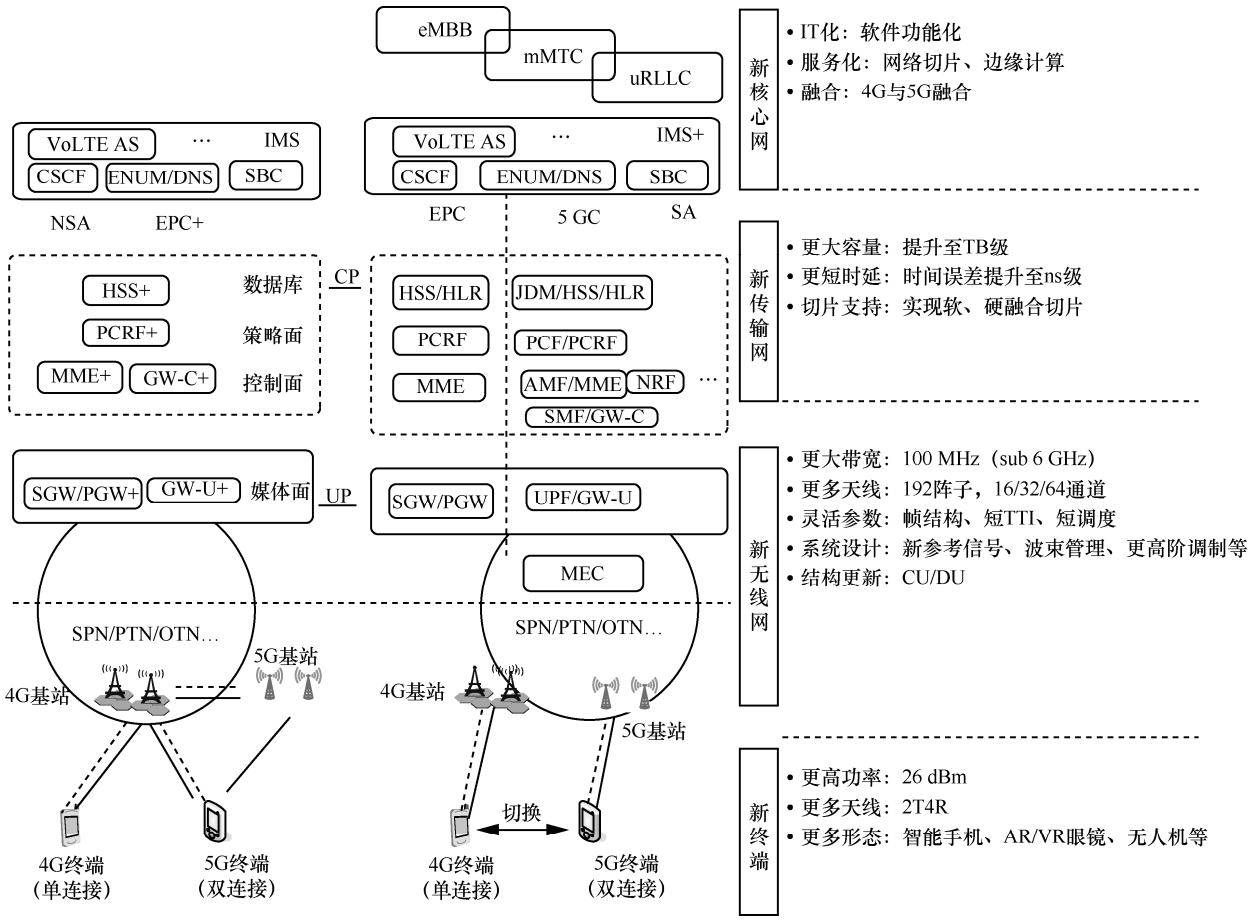


图 1　5G 端到端技术变革

新技术如何选择及准确引入。本文将以对 5G 架构选择与边缘计算（MEC）新技术引入两个关键问题的阐述为例。

如前文所述，5G 提供两种组网架构：NSA 非独立组网与 SA 独立组网。由于 NSA 组网架构中核心网沿用 4G 核心网，无法应用网络切片、控制面/用户面分离等 5G 核心网新的技术，因此 NSA 架构虽然能够规避 5G 核心网成熟较晚的时间风险，能够实现快速部署，但是会带来仅能满足初期 eMBB 大带宽高容量补充的业务需求；而 SA 作为全新的 5G 组网架构，能够实现 5G 网络全部功能，更好地支持 5G 的新业务与新特性，但是要承担 5G 端到端成熟的时间风险。两种组网架构的选择要综合频谱资源、无线覆盖能力、端到端成熟时间与业务发展匹配等因素。

- 选择方案一：直接部署 SA。商用初期即要同时满足大带宽和低时延两种业务需求，且与 5G 端到端成熟时间相匹配，此外，5G 频段划分和新技术应用能够满足连续组网需求，满足上述条件则选择直接部署 SA。
- 选择方案二：先部署 NSA，后续演进至 SA。商用初期的业务需求主要是大带宽业务，5G 端到端特别是核心网成熟较晚，5G 频段划分和新技术应用无法满足连续组网需求，则选择先部署 NSA，后续演进至 SA 方案。此时，LTE 无线网和 EPC 核心网均会面临二次改造。

MEC 是指核心网络和业务能力下沉，通过本地分流和预处理达到降低时延、节省带宽和提升用户体验的目的，属于典型的时延驱动型分布式架构。MEC 部署位置有多种选择，如图 2 所示，地市核心机房、骨干汇聚机房和基站机房，部署成本依次抬升，但是时延依次降低，因此，MEC 部署位置的选择脱离不了业务的应用需求，特别是垂直行业的业务需求。

NSA/SA 架构、MEC 部署的选择与引入策略除了理论分析以外，均需要通过规模技术试验进行实践与验证，其他新技术也是如此。因此，5G 规模试验设计了大量的测试样例用于支撑 5G 新技术引入策略的判断，是 5G 规模试验最为首要的内容之一。

2.2 建网方案

与以往 3G、4G 类似，建网方案的确定是网络商用面临的关键问题之一。以无线侧为例，建网方案需要考虑频谱问题、覆盖问题与容量问题。

频谱的选择对网络建设方案影响巨大，直接影响网络规模、无线网络质量、组网方案等。2017 年 11 月 9 日，工业和信息化部下发了《工业和信息化部关于第五代移动通信系统使用 3 300~3 600 MHz 和 4 800~5 000 MHz 频段相关事宜的通知》（工信部无[2017]276 号），

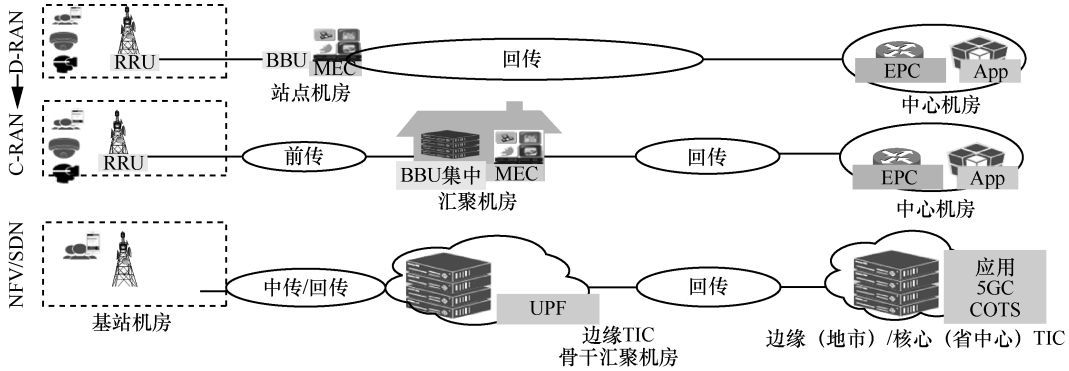


图 2 MEC 多种部署位置

确定了 6 GHz 以下 5G 可以使用的频率。由于全球 6 GHz 以下频谱聚焦 3.5 GHz 频段，4.9 GHz 设备成熟度较 3.5 GHz 晚 0.5~1 年左右，且与 4.9 GHz 相比，3.5 GHz 电波传播损耗较低、性能较好。从建网方案角度考虑为了更好地利用好这些频段资源，需要认真分析不同频段下网络性能。

网络覆盖问题中，电波传播特性的研究是核心。3.5 GHz/4.9 GHz 将带来更大的传播损耗，通过在济南、阳泉、杭州等多地的测试验证工作，发现在城区场景，3.5 GHz 传播损耗比 D 频段高 4~6 dB，4.9 GHz 较 3.5 GHz 传播损耗高 5~6 dB。考虑 64T64R 天线增益、发射功率、接收机灵敏度的影响，通过仿真可以得出，5G 的 3.5 GHz 覆盖能力与 D 频段相当，这一结论还需要在 5G 规模试验中进一步实践与验证。5G 网络覆盖问题中，由于上下行天线通道数差异性较大，由此带来的上下行覆盖差异的问题也是重点关注问题之一。如果不考虑任何覆盖增强技术，通过仿真可以看出，相同覆盖半径条件下，3.5 GHz 64 通道下行边缘速率为 55 Mbit/s，而上行边缘速率仅为 178 kbit/s。为了解决这个问题，5G 主要的引入技术有 SUL 与 CA，两者的本质均是为小区边缘用户提供了使用低频资源的能力，从而提高边缘用户的体验，但 CA 条件下下行方向使用低频资源，但 SUL 不行。上行业务的感知将会直接影响 5G 网络商用的节奏，因此在 5G 规模试验中对上行增强方案的验证将对 5G 网络上行业务感知的提升起到重要的促进作用。5G 不同频段的覆盖性能、上行增强技术带来的效果也是 5G 规模试验的重点测试内容之一。

大带宽、多天线技术的引入，使得 5G 与 4G 相比，容量性能将会有显著提升。同样通过理论仿真，可以初步看出 5G 小区下行/上行平均吞吐量为 4G 网络的 13 倍/20 倍，剔除单载波带宽 5 倍增益，频谱效率提升 2~4 倍；下行 4 流，上行双流时，5G 单用户上行峰值速率约为 285 Mbit/s，约为 4G 网络的 19 倍，下行峰值速率约为 1.5 Gbit/s，约为 4G 网络的 17 倍。16 通道以及 64 通道多天线技术与产品的选择除了影响网络性能，对于建网成本与建网难度都有影响，因此选择与业务场景匹配的方案将更为合理，在 5G 网络真正商用之前通过试验网进行评估能为技术选择提供科学技术依据。

2.3　业务需求

从上文的阐述中，可以看出，5G 网络面对的业务需求更加丰富多样。因此，在 5G 网络商用之前就要考虑与判断这张网络服务的对象是谁，这个对象对网络的需求具体是什么。经过初步研究与分析，5G 网络在起步阶段的主要业务是高清及超高清视频、AR/VR、云端游戏，属于 eMBB 场景范畴；发展阶段将会产生工业制造、自动驾驶、远程医疗和智慧交通等 uRLLC 场景的业务；后续到达成熟期，则会面临海量物联网的大规模应用与连接需求。本文将从中筛选出智能网联车、远程控制、AR/VR/高清视频、智能制造四大垂直行业，其对于网络的需求具体见表 1。

针对不同业务的应用示范，将在 12 个示范城市完成相关测试验证工作，测试工作将与 5G 规模试验工作同步进行。

3　5G 规模试验

3.1　原则与目标

5G 规模试验的总体目标是通过 5G 规模试验，验证关键技术与性能，支撑明确技术路线及引入策略，完成网络规划建设方案制定、摸索优化运营经验，推动端到端产业成熟，力争实现 5G 全方位引领，同时为运营商今后 5G 建设和运营培养储备人才。

5G 规模试验在杭州、广州、苏州、武汉、上海分别选择华为、中兴、爱立信、大唐、诺基亚

表 1　垂直行业对网络的要求

		带宽需求	时延需求	其他
智能网联车	V2V（车车通信）		<5 ms	可靠性 99.999%
	V2I（车路协同）		<10 ms	
	V2P（行人告警）		<10 ms	
	V2N（车上娱乐）	100 Mbit/s~1 Gbit/s		
远程控制		100 Mbit/s~1 Gbit/s	<20 ms	
AR/VR/高清视频	标准 4K	45 Mbit/s	<20 ms	
	2K VR	22 Mbit/s	<20 ms	
	4K VR	75 Mbit/s	<16 ms	
	10K VR	863 Mbit/s	<12 ms	
智能制造	实时控制	kbit/s	5~50 ms	
	工业穿戴	100 Mbit/s~1 Gbit/s	<100 ms	
	无线调度与定位	kbit/s	50 ms~1 s	
	传感/表计采集	<Mbit/s		连接：几百~几千平方米； 采集频次：次/秒~次/天

开展，逐步建成每城市百站规模试验环境，考虑测试结果的完备性，每城市至少测试两种品牌终端产品。每城市均划分无线网、核心网、传输网、终端 4 条测试线并行测试，提高效率。5G 规模试验遵循先内后外的原则，即先开展实验室测试，具备一定条件后，进行外场规模试验。

3.2　总体规划

5G 规模试验总体规划分为两个阶段，各有侧重。第一阶段主要是验证关键功能及性能验证，用于支撑技术路线决策，推动设备性能稳定，形成初步的端到端组网能力；第二阶段面向商用，进行网络规划、组网、优化、网管、运营、异厂商互通、网元融合等测试，全面达到商用水平，发展友好用户。考虑端到端产业进度，5G 规模试验第一阶段优先启动 NSA 组网架构相关测试，再进行 SA 组网架构相关测试。

3.3　试验内容

5G 规模试验内容分为无线网、核心网、传输网、终端四大类。目前，无线网和终端测试内容已经明确，分为基本性能对比、NSA 专项、多天线关键技术、室内外多频段、CU/DU 部署方案、5G 覆盖增强技术、终端测试 8 项内容。

其中，基本性能主要从覆盖、吞吐量、时延、可靠性等方面分别对 NSA 组网架构和 SA 组网架构条件下的 5G 网络性能进行评估，为 NSA 与 SA 的路线决策提供支撑；NSA 组网架构商用面临 LTE 锚点频段选择等问题，NSA 专项旨在验证 NSA 上述多种方案的基本性能，指导未来商用建设方案的选择，同时验证 LTE 与 5G 共存能力；多天线关键技术则是为了测试验证不同通道天线产品适用的场景、在不同场景下的多天线配置方案及其性能；室内外多频段则是为了通过不同频段组网方案的测试验证，为未来 5G 频率选择提供支撑；CU/DU 部署方案的目的在于验证 CU-DU 分离与合设的性能差异；5G 覆盖增强技术则是为了验证 SUL、CA 等增强技术对于 5G 覆盖、容量性能的提升效果、适用的场景及成本代价；终端测试的意义在于优化终端实现，推动终端产品成熟。

4　结束语

本文首先介绍了全球 5G 产业发展情况，然后

从新技术引入策略、建网方案和业务需求 3 个角度梳理了面向商用的 5G 关键问题，并对 NSA/SA 架构选择、MEC 部署、频率、覆盖与容量、垂直行业业务需求等典型问题进行分析；最后介绍了此次 5G 规模试验的基本情况，包括原则与目标、总体规划与试验内容。

参考文献：

[1] 3GPP. Base station (BS) radio transmission and reception (release 15): TS38.104 V15.1.0[S]. 2008.

[2] 3GPP. Physical channels and modulation (release 15): TS38.211V15.1.0[S]. 2008.

[3] 倪善金, 赵军辉. 5G 无线通信网络物理层关键技术[J]. 电信科学, 2015, 31(12): 48-53.
NI S J, ZHAO J H. Key technologies in physical layer of 5G wireless communications network[J]. Telecommunications Science, 2015, 31(12): 48-53.

[4] 张建敏, 谢伟良, 杨峰义, 等. 移动边缘计算技术及其本地分流方案[J]. 电信科学, 2016, 32(7): 132-139.
ZHANG J M, XIE W L,YANG F Y, et al. Mobile edge computing and application in traffic offloading[J]. Telecommunications Science, 2016, 32(7): 132-139.

[作者简介]

刘玮（1986−），女，中国移动通信集团设计院有限公司工程师，长期从事移动通信网络规划、优化等新技术研究与工具开发工作。

董江波（1978−），女，博士，中国移动通信集团设计院有限公司教授级高级工程师，长期从事移动通信网络规划、优化等新技术研究工作。

任冶冰（1987−），男，中国移动通信集团设计院有限公司工程师，长期从事移动通信网络规划、优化等新技术研究与相关通信软件研发工作。

专题：5G

面向 5G 新空口技术的 Polar 码标准化研究进展

谢德胜，柴蓉，黄蕾蕾，陈前斌
（重庆邮电大学移动通信重点实验室，重庆 400065）

摘　要：5G 的业务特性及能力要求为新空口（new radio，NR）设计更加高效的新型信道编码方案，极化码（Polar 码）因具备优异的性能已被确定为 5G 增强移动宽带（enhanced mobile broadband，eMBB）场景控制信道的编码方案。在对 5G 及 Polar 码进行概述的基础上，详细阐述面向 5G NR 的 Polar 码标准化工作主要研究内容，包括码构建、序列设计、速率匹配及信道交织等。继而针对第三代合作伙伴计划（3rd Generation Partnership Project，3GPP）各成员机构在无线接入网层一（Radio Access Network Layer 1，RAN1）标准化工作组提出的 Polar 码编码方案进行分析和对比，并在此基础上对 Polar 码标准化相关研究进展进行总结。

关键词：5G；极化码；码构建；序列设计；速率匹配
中图分类号：TN911.22
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018237

Standardization of 5G new radio technology oriented Polar code

XIE Desheng, CHAI Rong, HUANG Leilei, CHEN Qianbin
Chongqing Key Lab of Mobile Communications, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: A more efficient new channel coding scheme was required for NR by the service features and capability indices of 5G. Due to its excellent performance, Polar code was specified as the control channel coding scheme for 5G enhanced mobile broadband (eMBB) scenario. An overview of Polar code was presented and the main existing research issues of the standardization for new radio(NR) technology oriented Polar code, including code construction, sequence design, rate-matching and channel interleaving, etc, were elaborated. Then some specific analysis and comparisons of the Polar code schemes which were proposed by 3GPP member institutes in RAN1 conferences recently were presented. Finally, a summary of the standardization works of Polar code was given on the basis of the above discussions.

Key words: 5G, Polar code, code design, sequence design, rate-matching

收稿日期：2018−03−06；修回日期：2018−08−10
基金项目：面向 3GPP 标准化的 5G 无线传输技术研究与评估项目（No.MCM20160105）；5G 高速无缝接入技术方案与试验系统研发项目（No.2016ZX03001010-004）
Foundation Items: The Joint Scientific Research Fund of Ministry of Education and China Mobile (No.MCM20160105), National Science and Technology Specific Project of China (No.2016ZX03001010-004)

1　引言

随着移动通信技术的快速发展和智能终端的日趋成熟，4G 已难以有效支持未来移动互联网和物联网高速发展带来的移动数据流量的高速增长、海量的设备连接以及差异化新型业务需求。为满足用户日益增长的各类移动业务需求，5G 概念应运而生，并已成为学术界和信息产业界的重要研究课题之一[1]。

近年来，国内外多家通信组织及相关机构均积极致力于 5G 的研发工作。2012 年 9 月，欧盟启动了面向 5G 系统的 5GNOW（5th generation non-orthogonal waveforms for asynchronous signaling）研究课题，主要针对 5G 物理层波形技术开展研究[2]。2012 年 11 月，欧盟正式启动 METIS（mobile and wireless communications enablers for the twenty-twenty information society）研究项目，针对如何实现未来移动通信需求开展广泛研究[3]。此外，欧盟在 2014 年启动了规模更大的科研项目 5G-PPP（5G public-private partnership），旨在加速 5G 研究和创新。2013 年，中韩两国分别成立 IMT-2020（5G）推进组及 5G 技术论坛，以推进 5G 技术标准的研发[4]。2015 年，在 3GPP 服务工作组会议（SA1-Service）上确定的版本 14（Re1ease-14）标准工作项目对 5G 需求进行明确规范[5]。同年，国际电信联盟无线通信部（International Telecommunication Union-Radio Communications Sector，ITU-R）明确了未来 5G 三大典型应用场景[6]，分别为增强型移动宽带（enhanced mobile broadband，eMBB）场景、大规模机器类通信（massive machine type communication，mMTC）场景和超高可靠性低时延通信（ultra-reliable and low latency communication，uRLLC）场景。不同应用场景具有不同性能需求。eMBB 场景要求支持更高的传输速率（峰值速率：上行链路达到 10 Gbit/s，下行链路达到 20 Gbit/s）、更高的频谱效率（峰值频谱效率：上行链路达到 12 bit/(s·Hz)，下行链路达到 30 bit/(s·Hz)）等；mMTC 场景要求支持更大连接数密度（1×10^6 个连接/km^2）、更低能耗（终端电池使用寿命达到 15 年）；uRLLC 场景要求支持更低的时延（上下行链路时延 0.5 ms，即端到端时延低于 1 ms）、更高的可靠度（达到 99.999 9%，即 1 ms 内的误帧率低于 10^{-6}）、更低的错误平层等。

5G 中业务需求的多样性及各类业务场景的典型特性均给传统移动通信技术，特别是现有的信道编码技术带来新的困难及挑战。5G 三大典型应用场景对 5G NR 信道编码关键要求见表 1，而 4G 中采用的信道编码方案 Turbo 码因在可靠性（Turbo 码存在译码错误平层）、编译码复杂度、译码吞吐量和编码效率等方面难以有效满足 5G 场景下的各种性能要求。亟需为 5G 新空口（new radio，NR）设计更加先进高效的信道编码方案，以尽可能小的业务开销实现信息快速可靠传输。

目前，国内外研究机构已针对 5G 信道编码技术开展了大量研究，并已达成部分共识。Polar 码因其理论证明可达到香农极限，且具有可实用的线性复杂度编译码能力而受到业界重视，成为 5G NR 信道编码方案的强有力候选者。在 2016 年 11 月召开的 3GPP RAN1#87 次会议上确定 eMBB 场景

表 1　5G NR 信道编码关键要求

eMBB	mMTC	uRLLC
高吞吐量下具有好的错误性能	低吞吐量下具有好的错误性能	低/中吞吐量下具有非常好的错误性能
高能量效率及高芯片效率	易于实施	低编码/译码时延
低编码/译码时延	高能量效率	非常低的错误平层

的 5G 短码块信道编码方案采用 Polar 码作为控制信道编码方案。基于此，本文首先对 Polar 码的基本概念及原理进行概述，继而对近年来国内外研究机构针对 Polar 码开展的标准化研究工作进行分析综述。

2 Polar 码概述

2008 年，土耳其毕尔肯大学 Arikan 教授[7]在国际信息论（International Symposium on Information Theory，ISIT）会议上首次提出信道极化（channel polarization）的概念。2009 年，Arikan 教授在参考文献[8]中对信道极化进行更为详细的阐述，并基于信道极化思想提出一种新型信道编码方法，即 Polar 码。

由于在理论上可被严格证明在低译码复杂度下能够达到信道容量，Polar 码一经提出即受到学术界及业界的广泛关注，并针对 Polar 码相关理论及应用开展深入研究。参考文献[9]中，Arikan 分析了 Polar 码的极化现象，并给出 Polar 码在二元删除信道（binary erasure channel，BEC）中的具体构造方法以及编译码过程。考虑到 Arikan E 给出的 Polar 码构造方法仅适用于 BEC 信道，具有较大的局限性，Mori 和 Tanaka 等人[10-11]借鉴低密度奇偶校验（low-density parity-check，LDPC）码的构造方法，提出采用密度进化（density evolution，DE）方式构造 Polar 码，以适用于任意二进制离散无记忆信道（binary discrete memoryless channel，B-DMC）[10-11]。随后 Tal 在参考文献[12]中针对 DE 方法的复杂度进行了研究，给出了更为有效的构造方法。此外，也有研究考虑 Polar 码在几类常见的连续信道中的应用，如高斯信道、瑞利信道及中继信道等[13-16]。近年来，较多研究考虑 Polar 码在更为实际的通信信道场景，如多址接入信道[17]、存在窃听的通信网络[18]、量子信道[19]及多阶调制系统[20-21]中的应用。

另外，也有研究考虑基于 Polar 码的信源编码。在对具体 Polar 码信源编码方案进行研究的基础上，提出基于 Polar 码的联合信源信道编码理论[22]。

2.1 信道极化概念

Polar码是基于信道极化理论提出的一种新的线性分组码。信道极化是指以特定方式对任意 $N=2^n(n\geqslant 0)$个独立的 B-DMC 进行组合分裂，随着信道数目 N 的增加，子信道特性呈现两极分化的现象[23]。信道极化过程包括信道组合和信道分裂两个过程，以下分别进行简要介绍。

2.1.1 信道组合

将 B-DMC 用 $W:X\to Y$ 表示，其中，X 表示输入向量集合，Y 表示输出向量集合，转移概率表示为 $W(y|x),x\in X,y\in Y$。W^N 表示 N 个互相独立的信道 W 同时使用，对于 $W^N:X^N\to Y^N$ 信道的转移概率为 $W^N(y_1^N|x_1^N)=\prod_{i=1}^{N}W(y_i|x_i)$。

信道组合过程是指通过递归算法对 N 个独立的 B-DMC 信道 W 进行组合以得到组合信道 $W_N:X^N\to Y^N(N=2^n,n\geqslant 0)$，其中，$X^N$、$Y^N$分别表示信道 W_N的输入及输出向量集合。以下针对不同 N 值进行简要分析。

（1）若 $N=1$，可得 $W_1=W$，即不进行信道组合。

（2）若 $N=2$，即将两个独立的 B-DMC 信道 W_1 进行组合得到组合信道 $W_2:X^2\to Y^2$。具体组合方式如图 1 所示，其中，$u_1,u_2\in X$ 为信道 W_2 的输入向量，$y_1,y_2\in Y$ 为相应输出向量，$x_1=u_1\oplus u_2$、$x_2=u_2$ 分别为两个独立信道 W 的输入向量。信道 W_2 的转移概率如式（1）所示：

$$W_2\left(y_1,y_2|u_1,u_2\right)=W(y_1|u_1\oplus u_2)W\left(y_2|u_2\right) \quad (1)$$

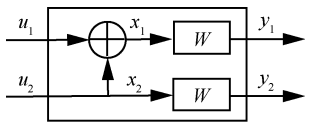


图 1 W_2 信道模型

（3）若 $N=4$，即将两个独立的组合信道 W_2 进行组合得到组合信道 W_4: $X^4\to Y^4$，如图 2 所

示，信道 W_4 的转移概率如式（2）所示：

$$W_4\left(y_1^4 \mid u_1^4\right)= W_2(y_1,y_2 \mid u_1 \oplus u_2,u_3 \oplus u_4)W_2\left(y_3,y_4 \mid u_2,u_4\right) \tag{2}$$

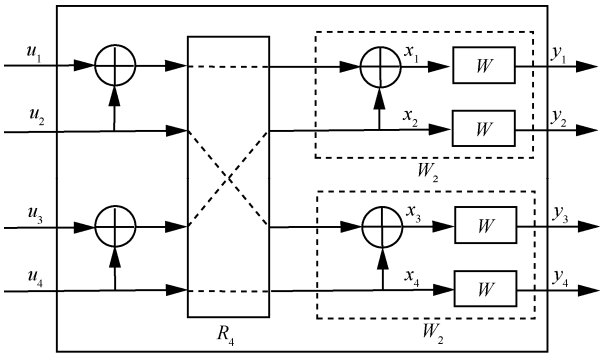


图 2　W_4 信道模型

通过图 2 可知，组合信道 W_4 的输入向量 u_1^4 至原始信道 W^4 的输入向量 x_1^4 的映射关系 $u_1^4 \to x_1^4$ 可表示为 $x_1^4 = u_1^4 G_4$，其中 $G_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$，因此，组合信道 W_4 与原始信道 W^4 之间的转移概率可表示为：

$$W_4\left(y_1^4 \mid u_1^4\right) = W^4(y_1^4 \mid u_1^4 G_4) \tag{3}$$

依此类推可得信道组合的一般形式，两个独立的信道 $W_{N/2}$ 可通过信道组合转化成信道 $W_N : X^N \to Y^N$。组合信道 W_N 的输入向量 u_1^N 到原始信道 W^N 的输入向量 x_1^N 之间的映射关系 $u_1^N \to x_1^N$ 可表示为 $x_1^N = u_1^N G_N$，其中 $G_N = B_N F^{\otimes n}$ 为 N 阶生成矩阵，B_N 为 N 阶比特反转矩阵，实现倒位功能，核心矩阵 $F=\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$，$F^{\otimes n}$ 为矩阵 F 的 n 阶克罗内克积，$F^{\otimes n} = F \otimes F^{\otimes n-1}$。由此可得到组合信道 W_N 和原始信道 W^N 的转移概率关系为：

$$W_N\left(y_1^N \mid u_1^N\right) = W^N(y_1^N \mid u_1^N G_N) \tag{4}$$

其中，$y_1^N \in Y^N, u_1^N \in X^N$。

2.1.2　信道分裂

信道分裂过程是将组合信道 W_N 分裂成 N 个二进制输入比特信道 $W_N^{(i)}$ 的过程。

若 $N = 2$，组合信道 W_2 分裂为 $W_2^{(1)}$ 及 $W_2^{(2)}$，即 $(W,W) \to (W_2^{(1)}, W_2^{(2)})$，对应的转移概率计算式为：

$$W_2^{(1)}(y_1^2,u_1) = \sum_{u_2} \frac{1}{2} W_2(y_1^2 \mid u_1^2) = \sum_{u_2} \frac{1}{2} W(y_1 \mid u_1 \oplus u_2) W(y_2 \mid u_2) \tag{5}$$

$$W_2^{(2)}(y_1^2,u_1 \mid u) = \frac{1}{2} W_2(y_1^2 \mid u_1^2) = \frac{1}{2} W(y_1 \mid u_1 \oplus u_2) W(y_2 \mid u_2) \tag{6}$$

对于任意组合信道 W_N，其分裂后的第 i 个信道 $W_N^{(i)}$ 对应的转移概率如式（7）所示：

$$W_N^{(i)}(y_1^N, u_1^{i-1} \mid u_i) = \sum_{u_{i+1}^N \in X^{N-i}} \frac{W(y_1^N, u_1^N)}{W(u_i)} = \sum_{u_{i+1}^N \in X^{N-i}} \frac{1}{2^{N-1}} W_N(y_1^N \mid u_1^N) \tag{7}$$

其中，y_1^N, u_1^{i-1} 表示信道 $W_N^{(i)}$ 的输出，u_i 为输入。

2.2　Polar 码编码

根据信道极化现象，可将原本相互独立的 N 个原始信道转化为 N 个信道容量不等的比特信道。当 N 趋于无穷大时，一部分信道的容量趋于 0，而另一部分信道的容量趋于 1。假设 K 个信道的容量趋于 1，N-K 个信道的容量趋于 0，可选择 K 个容量趋近于 1 的信道传输信息比特，选择 N-K 个容量趋近于 0 的信道传输冻结比特，即固定比特，从而实现由 K 个信息比特到 N 个编码比特的一一对应关系，也即实现码率为 K/N 的 Polar 码的编码过程。

Polar 码的具体编码方式可表示为：

$$x_1^N = u_1^N G_N \tag{8}$$

其中，$x_1^N = (x_1, x_2, x_3, ..., x_N)$ 为编码比特序列，$u_1^N = (u_1, u_2, u_3, ..., u_N)$ 为信息比特序列，$G_N = B_N F^{\otimes n}$ 为对应的 N 阶生成矩阵，B_N 为 N 阶比特反转矩阵，$F=\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$，$F^{\otimes n}$ 为矩阵 F 的 n 阶

克罗内克积，$F^{\otimes n}=F\otimes F^{\otimes n-1}$。Polar 码可由参数 (N,K,A,u_{A^c}) 的陪集 G_N 码定义[8]，其中 N 为码长，K 为信息比特个数，A 为信息比特位置集合，并且 A 中元素个数等于 K，u_{A^c} 为冻结比特所对应的序列，由于冻结比特所在的信道特性极差，在信息传输过程中一般固定设为 0。

由于上述编码中的生成矩阵 G_N 中存在比特反转矩阵 B_N，故该编码方式也称为比特反转编码。在 3GPP 中已确定 Polar 码采用无比特反转编码，并把采用该编码方式得到的 Polar 码称为“基本 Polar 码”，其生成矩阵为 $G_N=F^{\otimes n}$ [24]。

2.3 Polar 码译码

将信息序列 u_1^N 编成码字 x_1^N 后经由信道 W^N 传输，接收端收到 y_1^N。接收端译码的过程就是根据已知的接收信号 y_1^N 得到信息序列 u_1^N 的估值 $\hat{u}_1^N$。典型 Polar 码译码算法为连续消除（successive cancellation，SC）译码算法，其基本思想是按序号从小到大的顺序依次对信息比特进行基于似然比的硬判决译码。

令 A 表示信息比特位置集合，A 的补集 A^c 表示冻结比特位置集合，SC 译码式如式（9）所示：

$$\hat{u}_i=\begin{cases}h_i\left(y_1^N,\hat{u}_1^{i-1}\right), & i\in A\\ u_i, & i\in A^C\end{cases} \tag{9}$$

其中，$h_i(y_1^N,\hat{u}_1^{i-1})$ 为信息比特的译码准则，定义如下：当 $i\in A^C$ 时，表明 $\hat{u}_i$ 为收发双方事先约定的比特，可直接判决为 $\hat{u}_i=u_i$；当 $i\in A$ 时，表明 $\hat{u}_i$ 为承载信息的信息比特，其判决要根据已经判决出来的 $\hat{u}_1^{i-1}$，然后计算其似然比（likelihood ratio，LR），如式（10）所示：

$$h_i\left(y_1^N,\hat{u}_1^{i-1}\right)=\begin{cases}0, & L_N^{(i)}\left(y_1^N,\hat{u}_1^{i-1}\right)\geqslant 1\\ 1, & \text{其他}\end{cases} \tag{10}$$

其中，$L_N^{(i)}(y_1^N,\hat{u}_1^{i-1})$ 为 LR，如式（11）所示：

$$L_N^{(i)}\left(y_1^N,\hat{u}_1^{i-1}\right)=\frac{W_N^{(i)}(y_1^N,\hat{u}_1^{i-1}|0)}{W_N^{(i)}(y_1^N,\hat{u}_1^{i-1}|1)} \tag{11}$$

当 Polar 码码长趋于无穷时，由于各个分裂信道接近完全极化，采用 SC 译码算法可确保对每个信息比特实现正确译码，从而可以在理论上使得 Polar 码达到信道的对称容量 $I(W)$。然而，对于较短或有限码长的 Polar 码，SC 译码算法性能并不理想。为提升有限码长 Polar 码性能，已有研究人员提出多个高性能译码算法，如置信传播（belief propagation，BP）译码算法[25]、线性规划（linear programing，LP）译码算法[26]、串行抵消列表（successive cancellation list，SCL）译码算法[27-28]、串行抵消堆栈（successive cancellation stack，SCS）译码算法[29]、混合串行抵消（successive cancellation hybrid，SCH）译码算法[30]等。

近年来，国内外多家公司针对 5G NR 场景下的 Polar 码标准化工作开展深入研究，并已取得阶段性成果。现有的面向 5G NR Polar 码标准化研究工作主要涉及码构建、序列设计、速率匹配和交织器设计等方面，将在后续各节逐一进行介绍。

3 Polar 码构建标准研究进展

3.1 Polar 码构建概述

Polar 码构建的关键是编码结构的设计。为提升译码性能、减少译码复杂度及控制信道盲检测的次数，3GPP 建议采用基于循环冗余校验（cyclic redundancy check，CRC）辅助 Polar 码方案进行码构建，具体编码结构为“*J*+*J'*+基本 Polar 码”，其中，*J* 表示 24 位 CRC 比特，主要用于错误检测及辅助译码；*J'*表示额外的 CRC/奇偶比特，主要用于辅助译码，针对不同物理信道可采用不同的值。CRC 辅助（CRC assisted，CA）Polar 码的编码和译码流程如图 3 所示。

3.2 Polar 码构建标准提案

目前，各公司所提的 Polar 码构建标准提案主要针对基于 CA Polar 码方案，具体包括（*J*+*J'*）位 CRC 比特分布方式选择以及 *J'* 比特长度选择等问题。

3.2.1 编码结构

综合考虑误块率（block error rate，BLER）、

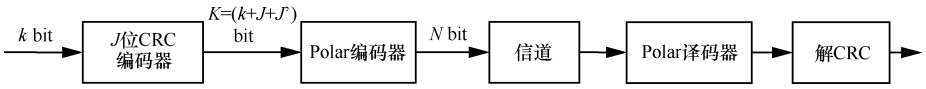
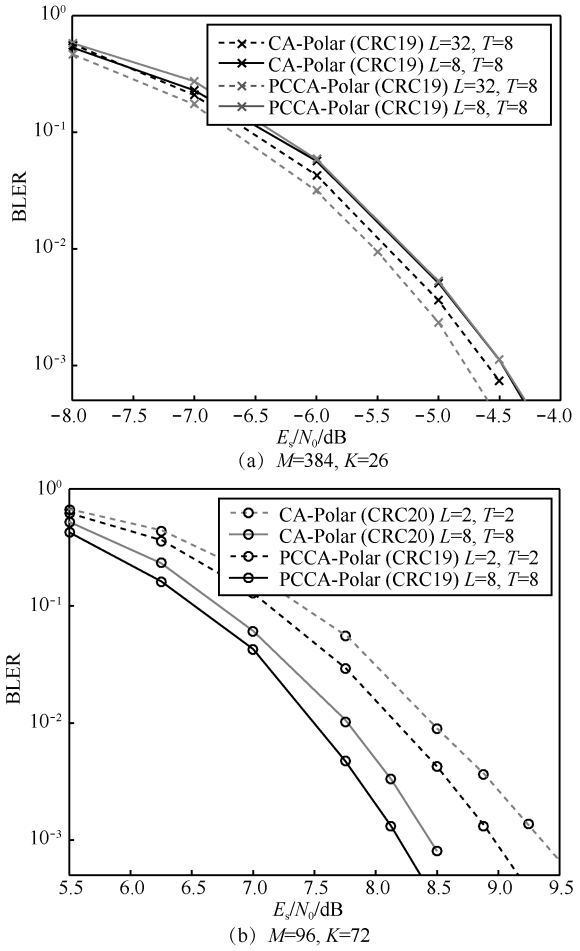


图 3　CRC 辅助 Polar 码编码和译码流程

虚警率（false alarm rate，FAR）及提前终止（early termination，ET）增益等因素，研究人员对 Polar 码构建方案提出具体建议。参考文献[31]中，华为技术有限公司（以下简称华为公司）提出基于奇偶校验 CRC CA Polar（parity check-CRC assisted，PC-CA Polar）码设计方案，该方案采用 3 bit PC 比特且 CRC 比特采用分布式方式以最大化辅助比特增益。提案中也对具有不同 CRC 长度的 PC-CA Polar 码及 CA Polar 码的 BLER 性能进行了分析对比，并建议在 NR 控制信道中采用 PC-CA Polar 码方案，如图 4 所示。



（a）M=384, K=26

（b）M=96, K=72

图 4　不同 CRC 长度的 PC-CA Polar 码及 CA Polar 码的 BLER 性能

大唐电信科技股份有限公司（以下简称大唐公司）在参考文献[32]中提出两类 Polar 码方案。其中，方案一采用 X 编码器，如 Hash 编码器或 CRC 编码器，对 J 位 CRC 编码器输出的编码向量进行编码，并将编码得到的附加比特向量添加到整个编码向量末端，将其作为 Polar 码编码器的输入。方案二首先将 J 位 CRC 编码器输出的编码向量分段，进而分别输入分段编码器（Hash 编码器或 CRC 编码器）进行编码，并将编码得到的附加比特向量添加至每个分段的编码向量末端，将其作为 Polar 码编码器的输入。由于方案一将 CRC 和附加比特向量添加在编码向量末端，采用该方案得到的码字在译码时不能获得 ET 增益。而方案二将 CRC 和附加比特向量分散在编码向量的不同位置，采用该方案得到的码字在译码时可获得 ET 增益。上述方案的编码结构如图 5 及图 6 所示。

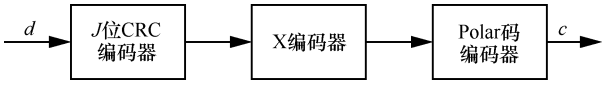


图 5　X-级联 Polar 码编码结构

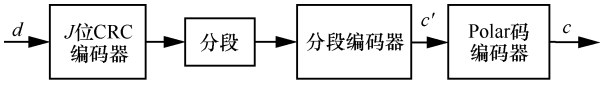


图 6　支持 ET 的分段 Polar 码编码结构

参考文献[33]中，日电（中国）有限公司（以下简称日电公司）研究在译码失败的情况下，支持 ET 译码的辅助比特 Polar 码构建，其中，辅助比特可采用 PC 比特、CRC 比特和 Hash 比特，并对分布式 CRC（distributed CRC，DCRC）辅助 Polar 码和 PC-DCRC 辅助 Polar 码进行 ET 性能评估，如图 7 所示。基于性能评估结果，日电公司建议采用 PC-DCRC 方案作为 Polar 码 ET 基准方案。

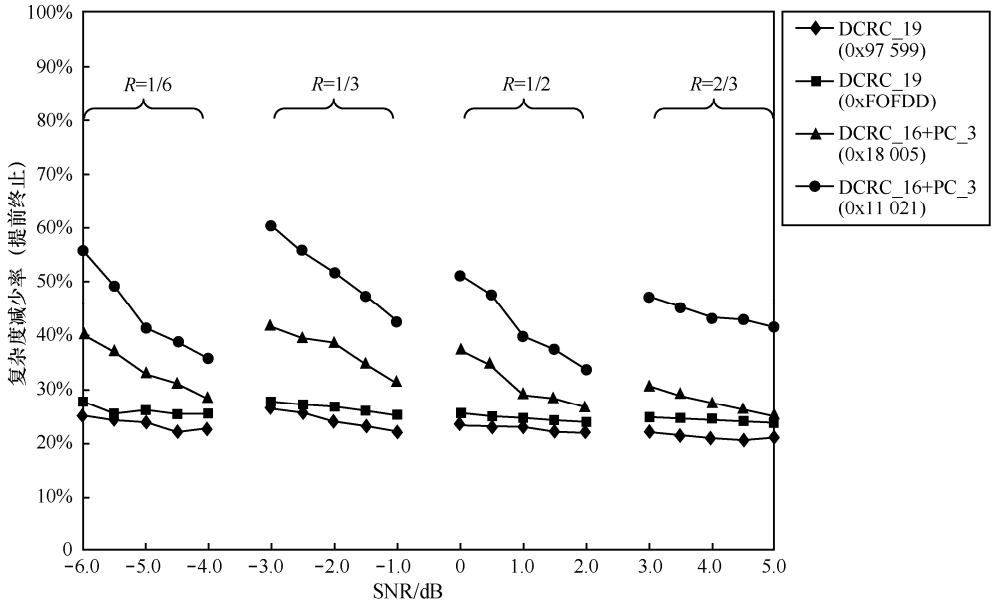


图 7　不同 ET 方案的 ET 性能评估

3.2.2　CRC 比特分布方式

在基于 CA-Polar 码构建阶段，（J+J'）位比特可采用分布式及非分布式两种方式分布。其中，分布式指将（J+J'）位比特以分布式的方式插入信息比特中，译码端只译出一部分比特即可进行一次 CRC 校验。非分布式指将（J+J'）位比特统一放置于信息比特后端，仅当译码端译出全部比特才可对该码字进行 CRC 校验。（J+J'）比特分布方式的选择也同样需要考虑 BLER 和 ET 增益等因素，研究人员针对（J+J'）比特分布方式的选择问题提出相关方案。

爱立信公司针对 DCRC Polar 码设计方案在参考文献[34]中给出了两种变体方法：非迭代 DCRC 方案和迭代 DCRC 方案。在非迭代 DCRC 方案中，DCRC Polar 码的辅助 CRC 比特采用非递归方式生成，即第 i 个辅助 CRC 比特的计算与第 i 个之前产生的辅助 CRC 比特无关，该方案的编码器结构如图 8 所示。在迭代 DCRC 方案中，DCRC Polar 码的辅助 CRC 比特采用递归方式生成，该方案编码器结构在图 6 所示的编码器结构上引入反馈机制，即第 i 个辅助 CRC 比特的生成需将信息比特和第 i 个之前产生的辅助 CRC 比特

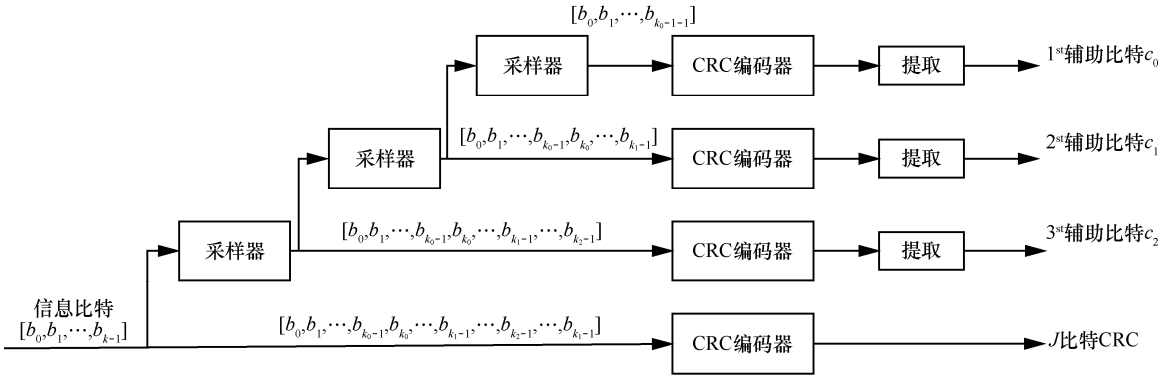


图 8　非迭代 DCRC 编码器结构

进行组合采样，继而送入第 i 个 CRC 编码器进行计算提取得到第 i 个辅助 CRC 比特。

三星电子公司（以下简称三星公司）在参考文献[35]中对各家公司提出的辅助比特 Polar 码构建方案进行概括，其中，与 ET 特性相关的码构建方案包括 DCRC 方案、分布式 PC 方案和多 CRC 方案；与 BLER 性能改善相关的码构建方案包括 CA 方案和 PC-CA 方案等，并建议在 NR 控制信道采用附加辅助比特数量 $J' = 3$ 的 Polar 码。

4　Polar 码序列设计标准研究进展

4.1　Polar 码序列设计概述

Polar 码序列设计问题即 Polar 码中信息比特信道和冻结比特信道的选择问题。由于 Polar 码基于信道极化理论构建，在进行 Polar 码编码时，首先需区分经信道组合和信道分裂后得到的 N 个比特信道的可靠程度，即区分可靠子信道（无噪声子信道）及不可靠子信道（全噪声子信道），进而将信息比特和检错比特放置在可靠子信道进行传输，将冻结比特放置在不可靠子信道进行传输。因此，如何度量各子信道可靠性成为 Polar 码序列设计的关键。

常见的子信道可靠性度量方法包括巴氏参数（Bhattacharyya parameter，BP）法[8]、DE 法[11]和高斯近似（Gaussian approximation，GA）法[36]等，其中，BP 法在 BEC 下采用递归方法计算信道的可靠性并且复杂度低，但在其他信道下，只能得到近似可靠性度量。DE 法通过跟踪各子信道概率密度函数（probability density function，PDF），对子信道错误传输概率进行评估。该方法适用于所有类型的 B-DMC，但具有较高计算复杂度。GA 法针对二进制输入加性高斯白噪声（additive white Gaussian noise，AWGN）信道，将 DE 法中的对数似然比（log-likelihood ratio，LLR）的 PDF 以高斯分布近似，从而简化子信道评估的复杂度，降低计算量。

4.2　Polar 码序列设计标准提案

对 Polar 码进行序列设计时，需综合考虑性能、信息粒度、与速率匹配的兼容性、复杂度及时延等因素，多家公司针对子信道可靠性度量方法开展了相关研究，并提出了相应的 Polar 码序列设计具体方案。

华为公司在参考文献[37]中提出 β 扩展算法。该算法基于极化权重（polarization weight，PW）对子信道进行排序，以度量子信道可靠性。该算法与 GA 性能接近，但具有较低复杂度。极化权重定义为 $f^{\mathrm{PW}}: x \to \sum_{i=1}^{n} b_i \beta^i$，其中，$x$ 表示比特信道索引值，b_i 为索引值的 n 比特二进制扩展集合 $B = (b_{n-1},\cdots,b_1,b_0)$中第 i 位比特，β 值的大小决定子信道的顺序，华为公司建议 $\beta=2^{1/4}$[37]。

Polar码中指示信息比特与冻结比特位置的顺序序列具有 Polar 码所固有的嵌套码构造特性，而单序列方案正是 Polar 码嵌套码结构特性的自然结果。华为公司在参考文献[38]中将单序列方案中的 PW 序列跟其他单序列在描述复杂度、前向兼容性、硬件实现复杂度方面进行比较，结果表明 PW 序列在以上 3 方面都具有优异的性能和额外的增益，因此建议在 NR 控制信道中采用单序列中的 PW 序列。

现有 Polar 码序列设计大多采用基于高斯均值近似的 DE 法，美国高通公司（以下简称高通公司）提出采用互信息（mutual information，MI）的 DE 法实现 Polar 码序列设计，即互信息密度进化（mutual information density evolution，MI-DE）法。在对 Polar 码的嵌套码构造特性进行深入分析的基础上，给出了基于 MI-DE 长序列构建方法，并对所提 Polar 码构建方法及爱立信、华为、乐喜金星（以下简称乐金）、台湾联发科技股份有限公司（以下简称联发科）、高通、三星、中兴通讯股份有限公司（以下简称中兴）7 家公司所提方法进行性能评估[39-41]。

参考文献[42]中，三星公司提出基于组合嵌套（combined-and-nested，CN）的 Polar 码序列设计方法，以实现对已有 Polar 码短序列的有效扩展。图 9 为 CN Polar 码序列设计概念图，由于母码长度小于或等于 32 bit 的 Polar 码序列是确定的且遵循偏序规则，故 Polar 码序列设计从母码长度大于 32 bit 开始。可基于如下方式得到母码长度为 64 bit 的 Polar 码序列：首先基于 DE 构建(N,K)分别为(64,1),(64,2),...,(64,63)的 Polar 码，其中，N 表示 Polar 码母码长度，K 表示不包括 CRC 比特的信息比特数；再将上述(N,K) Polar 码组合即可得到母码长度为 64 bit 的速率兼容 Polar 码序列。基于母码长度为 64 bit 的 Polar 码序列，可采用嵌入方式得到母码长度为 128 bit 的 Polar 码序列，通过该方式，最终可得到母码长度为 1 024 bit 且速率与长度兼容的 CN 序列。

图 10 为华为公司所提 PW 序列分别与高通公司所提 MI-DE 序列以及三星公司所提 CN 序列在不同信息块长度下的性能比较。

5 Polar 码速率匹配标准研究进展

5.1 Polar 码速率匹配概述

通信系统的速率匹配是指信道编码后的比特流速率应与信道传输速率相一致。由于在不同时间间隔内，传输信道的数据量大小是动态变化的，而所配置的物理信道时频资源则保持固定不变，因此，需要对输入比特流进行调整从而使其符合物理信道的承载能力。

根据信道编码后的输出比特流与物理信道承载能力的关系，可采用重复、打孔和缩短操作来实现速率匹配。若传输的编码比特数大于母码长度，采用重复编码，即对某些比特位进行重复；若传输的编码比特数小于母码长度且码率小于或等于最优码率阈值，采用打孔编码，即对某些比特位进行打孔，其相应的 LLR 在接收端设置为 0；若传输的编码比特数小于母码长度且码率大于最优码率阈值，可采用缩短编码，即对某些比特位进行删除，其相应的 LLR 在接收端设置为较大的值。

Polar码速率匹配相关研究涉及速率匹配具体

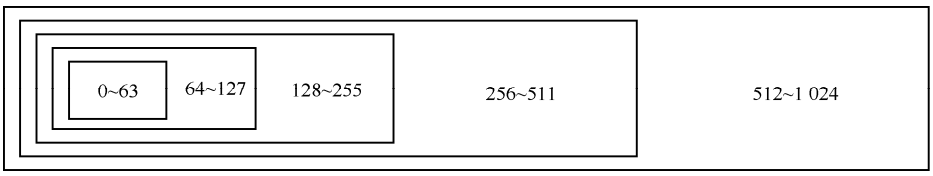
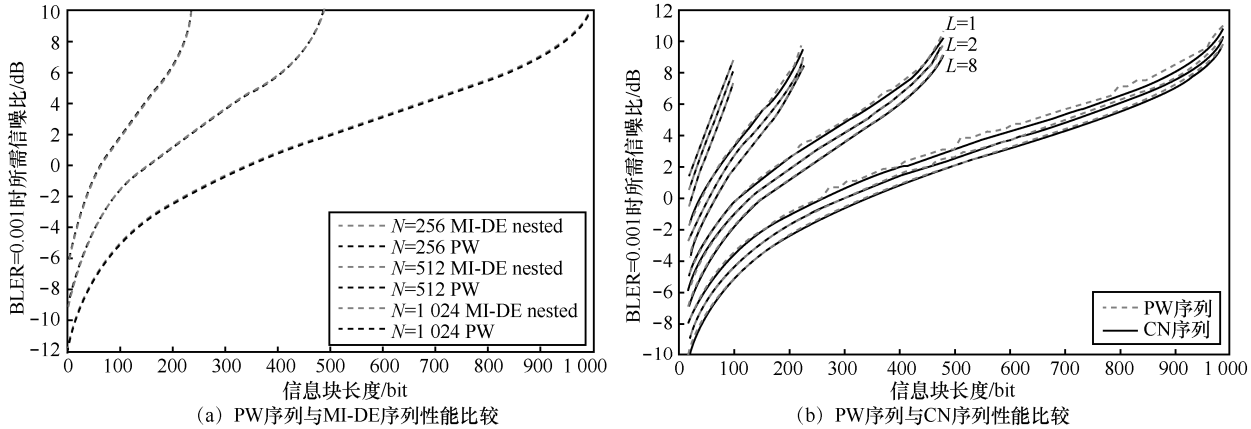


图 9　组合嵌套序列设计概念



（a）PW序列与MI-DE序列性能比较　　（b）PW序列与CN序列性能比较

图 10　各序列 BLER 性能比较曲线

方案的选择以及执行重复、打孔和缩短操作时循环缓冲器起始位置选择等相关问题。

5.2　Polar 码速率匹配标准提案

综合考虑需灵活支持 Polar 码不同码长、BLER 性能及实施复杂度等因素，研究人员对 Polar 码速率匹配提出相关方案和建议。参考文献[43]中，联发科公司提出具有统一模式的循环缓存器速率匹配方案。在该方案中，Polar 码编码器输出的编码比特被划分为 B_0、B_1、B_2 和 B_3 共 4 部分，循环缓存器由 3 部分组成，其第一部分为 B_0，第二部分由 B_1 和 B_2 隔行交织得到，第三部分为 B_3，通过读取缓冲区中不同长度的编码比特从而实现重复、打孔和缩短。参考文献[44]中，联发科公司对具有统一模式的循环缓存器速率匹配方案在时延、实现复杂度和性能增益方面进行了分析。考虑到该方案中使用的中间隔行交织操作简单，易于集成到编码器的输出功能和解码器输入功能中，且能够在信道比特交织过程中实现简单的并行块交织，从而获得最低的总体复杂度和时延，建议 NR Polar 码编码链中采用具有统一模式的循环缓存器速率匹配方案。

参考文献[45]中，华为公司提出基于分组的速率匹配方案，该方案首先将母码长度为 N 的编码比特划分成 32 个等长的群组，进而以群组方式完成缩短或打孔操作。该速率匹配方案可实现性能和复杂度之间较好的平衡。参考文献[46]中，中兴公司提出二维循环缓冲器速率匹配方案。该方案中，采用与 LTE 中子块交织器类似的行列交织器实现二维循环缓冲器，具体方式为将 Polar 码编码器输出的编码比特从上至下逐行写入交织器，继而执行行交织操作。在执行速率匹配时，编码比特按行被打孔和缩短且以自然顺序选择每行中被打孔或缩短的比特。

高通公司在参考文献[47]中提出了基于长序列信息比特分配调整的块速率匹配设计方案，该速率匹配方案包括 3 部分：确定冻结比特打孔/缩短位置、码参数确定及信息比特分配调整、比特选择。三星公司在参考文献[48]中提出基于子块排序的 Polar 码短序列速率匹配方案，该方案首先将编码器输出的码字向量划分为 16 个子块并将这些子块按特定顺序排序，进而将排序后的子块码字比特存储至循环缓存器，相应的速率匹配操作只需按顺序从循环缓存器中提取码字比特即可实现。图 11 为三星公司（SS）、联发科公司（MTK）及中兴公司（ZTE）所提速率匹配方案在不同信息块大小下的性能比较。

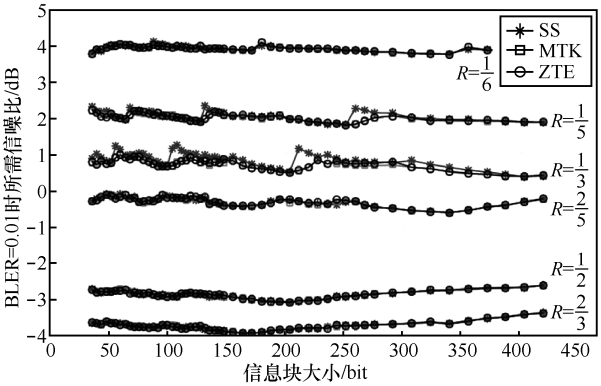


图 11　BLER=0.01 时不同速率匹配方案所需信噪比与信息块大小曲线

6　Polar 码信道交织标准研究进展

6.1　Polar 码信道交织概述

信道交织技术是实际移动通信环境下改善信号衰落的一种通信技术。该技术通过对比特进行分散化处理，可将一条消息中的相邻比特以非相邻的方式发送。在信息传输的过程中即使发生成串差错，在恢复成相邻比特串消息时，成串差错将转换为单个错误比特，因此接收端可采用纠正随机差错的编码技术实现纠错以恢复原消息。

根据进行交织操作的对象为数据块或比特，交织操作分别对应子块交织及比特交织。子块交织是指将编码后的码字分散为多路信息比特流送入速率匹配的子块交织器中，子块交织

器分别对 3 路信息比特流进行交织，以打乱信息比特的顺序，使噪声随机分布，降低出错概率。比特交织是指将比特流中的比特重新排列，从而使差错随机化的过程，也即将比特流进行分组，相继取出分组中的各比特，组成新的比特分组。

各公司对 Polar 码交织问题的相关研究包括编码及速率匹配之后信道交织方案选择以及信道交织器设计等。

6.2 Polar 码信道交织标准提案

考虑到 Polar 码性能对信道质量非常敏感，特别是在应用高阶调制和衰落信道时，因此需采用信道交织来提高性能。但在考虑信道交织器的类型及信道交织器在编码链中的位置时，需综合考虑 BLER 性能、时延及实现复杂度等因素。鉴于此，多家公司提出了相应的信道交织方案。

华为公司在参考文献[49]中对 Polar 码编码链研究表明：在调制阶数高于正交相移键控（quadrature phase shift keying，QPSK）的情况下，当调制符号失真时，通常会出现突发错误。为获得更好的编码性能，可使用比特交织编码调制（bit-interleaved-coded-modulation，BICM）方案使得编码比特随机化，将突发错误分散至码字的离散位置上，从而使解码尽可能成功。参考文献[50]给出了华为公司设计的信道交织器，该信道交织器由比特收集器和行列交织器组成，如图 12 所示。比特收集器根据编码比特索引集合和信息比特索引集合将比特序列 $x_0,x_1,...,x_{M-1}$ 划分为 f_m 和 g_n 两部分，继而级联 f_m 和 g_n 得到序列 y。行列交织器则将序列 y 按行写入，继而按列读取得到序列 e。

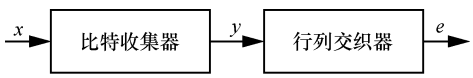


图 12　信道交织器

多家公司也针对下行链路和上行链路中的 Polar 码信道交织器结构设计问题分别进行讨论。参考文献[51]中，爱立信公司针对下行链路传输提出并行矩形信道交织器方案。该方案首先将长度为 N 的序列划分为长度为 N_1 和 N_2 的两段，进而采用交织深度为 5 和 11 的两个并行矩形交织器分别对 N_1 和 N_2 进行交织，最后交替输出两个并行矩形交织器中的比特以形成最终交织序列。该并行矩形交织器具有规则的结构且支持并行存储访问，因此较容易实现且处理时延较低。

大唐公司在参考文献[52]中也提出基于并行矩形信道交织器的交织方案，如图 13 所示。该方案由 4 个步骤组成：串/并转换，将长度为 M 的编码比特序列转换为长度为 m 的 31 段子序列，即完成简单的串/并转换；分组，将 31 段子序列分为 L 组，用 g 表示组号，x 表示段号，分组规则可表示为 $g = x \bmod L$；并行交织，使用不同交织深度的行列子交织器对不同组中的比特进行随机化；并/串转换，将行列子交织器的输出序列进行并/串转换得到最终序列。

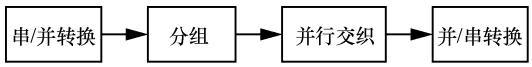


图 13　交织器模型

美国交互数字公司在参考文献[53]中给出了现有 CRC 生成多项式和交织器模式的 FAR 评估结果。基于评估结果，美国交互数字公司给出了现有交织器模式的修改方案以获得更好的 FAR 性能。参考文献[54]中，爱立信公司针对 NR 物理下行控制信道（physical downlink control channel，PDCCH）中的信道交织器方案进行研究，并分别给出了并行矩形信道交织器、矩形信道交织器及不使用信道交织器 3 种方案在不同控制信道单元（control channel element，CCE）聚合等级（aggregation level，AL）及资源粒子组（resource element group，REG）下的性能评估结果。

诺基亚公司提出在下行控制信道编码链中已

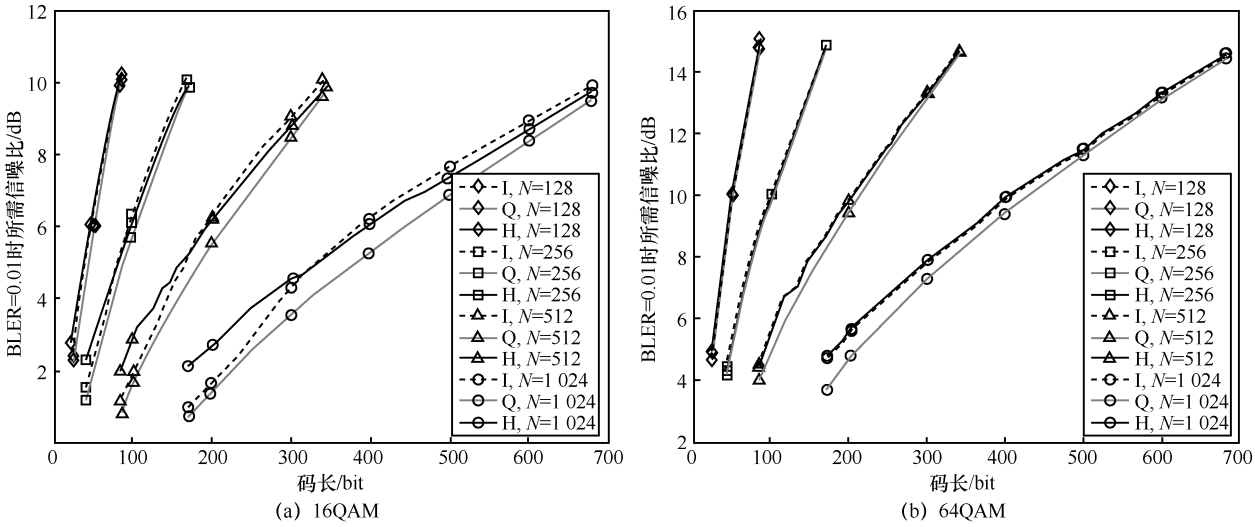
采用的 3 个交织器外引入第 4 个交织器[55]。其中，现有 3 个交织器分别为：位于编码阶段的交织器，实现将 CRC 比特以分散的方式插入信息比特中；位于速率匹配阶段的子块交织器，实现将 N 位编码比特划分为 32 个子块，并对 32 个子块重新排序；位于速率匹配阶段的比特交织器，对循环缓存器中的序列进行交织处理得到实际传输的编码序列。在使用上述 3 个交织器的基础上，诺基亚公司提出在下行控制信道编码链中引入比特级信道交织器，作为第 4 个交织器，从而实现 BLER 性能提升。在参考文献[56]中高通公司给出其所提的交织器设计方案（图 14 中 Q 表示），并对华为公司（图 14 中用 H 表示）所提方案、美国交互数字公司（图 14 中用 I 表示）所提方案和其所提方案在 AWGN 信道下进行性能评估，如图 14 所示。

针对上行控制信道交织器结构设计问题，诺基亚公司在参考文献[57]中使用最小扩展距离和平均扩展距离作为信道交织器设计准则，并提出一种具有不同读/写方式的三角形交织器。该交织器消除了恒定最小扩展距离的限制，使 BLER 性能显著提升，且支持高效的读写操作使端到端读写时延显著降低。

7　结束语

Polar 码虽起步晚，但因其优异的理论基础已被确定为 5G eMBB 场景的控制信道编码方案。目前在 3GPP RAN1#87 次会议及其后续会议讨论和研究的主要内容集中在短码的设计及实现上，如与 Polar 码相关的码构建、序列设计、速率匹配以及信道交织等问题。相应解决方案已在 3GPP RAN1 各次会议上达成，其相应的性能也能满足 eMBB 场景控制信道性能需求，但 Polar 码在 5G 的实际应用中仍有待进一步讨论和研究。

考虑到当前关于 Polar 码的相关标准主要是针对 eMBB 场景下的短码方案，在 5G 移动通信的新型场景 mMTC 和 uRLLC 中，采用何种信道编码方案（LDPC 码、Polar 码、Turbo 码）还需进一步讨论和研究，但可预见的发展趋势是多种信道编码方案配合使用；另外，即使采用何种编码方式已经确定，但如何根据具体业务需求灵活选择适合的编码方案和编码参数以及相应的性能及复杂度评估，也是 5G 后续工作面临的一个重要问题；最后，上述编码方案在实际通信系统的应用研究，如编译码器的设计、与 HARQ 和调制解调的联合设计、在多天线传输方案中的应用及硬



(a) 16QAM　　(b) 64QAM

图 14　不同调制阶数下各公司所提交织器性能曲线

件实现技术等，也是 5G 信道编码应用时的重点研究内容。

本文在对 Polar 码进行概述的基础上，对 Polar 码原理进行了阐述，进而对 3GPP RAN1 各次会议中各公司针对 Polar 码的标准化工作进行探讨，并从 Polar 码构建、序列设计、速率匹配以及信道交织等多方面进行总结，以期为后续 5G NR 信道编码研究工作和理论学习提供参考。

参考文献：

[1] 杨峰义, 张建敏, 王海宁, 等. 5G 网络架构[M]. 北京: 电子工业出版社, 2017.
YANG F Y, ZHANG J M, WANG H N, et al. 5G network architecture[M]. Beijing: Publishing House of Electronics Industry, 2017.

[2] 5GNOW. 5th generation non-orthogonal waveforms for asynchronous signaling[R]. 2012.

[3] METIS. Mobile and wireless communications enablers for the twenty-twenty information society. EU 7th framework Programme project[R]. 2012.

[4] 任永刚, 张亮. 第五代移动通信系统展望[J]. 信息通信, 2014(8): 255-256.
REN Y G, ZHANG L. Prospect of the fifth-generation mobile communication system[J]. Information and Communications, 2014(8): 255-256.

[5] 3GPP. Minutes of 3GPP TSG SA WG1 meetings#69[R]. 2015.

[6] IMT. White paper on 5G vision and requirements_V1.0[R]. 2014.

[7] ARIKAN E. Channel polarization: a method for constructing capacity-achieving codes[C]//IEEE International Symposium on Information Theory, July 10-15, 2008, Barcelona, Spain. Piscataway: IEEE Press, 2008: 1173-1177.

[8] ARIKAN E. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels[J]. IEEE Transactions on Information Theory, 2009, 55(7): 3051-3073.

[9] KORADA S B, SASOGLU E, URBANKE R. Polar codes: characterization of exponent, bounds, and constructions[J]. IEEE Transactions on Information Theory, 2010, 56(12): 6253-6264.

[10] MORI R, TANAKA T. Performance and construction of polar codes on symmetric binary-input memoryless chan-nels[C]// IEEE International Workshop on Statistical Physics and Computer Sciences, June 28-July 3, 2009, Seoul, Korea. Piscataway: IEEE Press, 2009: 1496-1500.

[11] MORI R, TANAKA T. Performance of polar codes with the construction using density evolution[J]. IEEE Communications Letters, 2009, 13(7): 519-521.

[12] TAL I, VARDY A. How to construct polar codes[J]. IEEE Transactions on Information Theory, 2013, 59(10): 6562-6582.

[13] ABBE E, BARRON A. Polar coding schemes for the AWGN channel[C]//IEEE International Symposium on Information Theory, July 31-August 5, 2011, St.Petersburg, Russia. Piscataway: IEEE Press, 2011: 194-198.

[14] ZHAO S, SHI P, WANG B. Designs of bhattacharyya parameter in the construction of Polar codes[C]//IEEE International Conference on Wireless Communications, Networking and Mobile Computing, September 24-26, 2011, Beijing, China. Piscataway: IEEE Press, 2011: 2-4.

[15] ZHAO S, SHI P, WANG B. Polar codes and its application in speech communication[C]//IEEE International Conference on Wireless Communication and Signal Processing, August 5-8, 2011, New York, USA. Piscataway: IEEE Press, 2011: 1-4.

[16] BLASCO-SERRANO R, THOBABEN R, ANDERSSON M, et al. Polar codes for cooperative relaying[J]. IEEE Transactions on Communications, 2012, 60(11): 3263-3273.

[17] ABBE E, TELATAR E. MAC polar codes and matro-ids[C]// IEEE Information Theory and Applications Workshop, January 6-8, 2010, Cairo, Egypt. Piscataway: IEEE Press, 2010: 1-8.

[18] MAHDAVIFAR H, VARDY A. Achieving the secrecy capacity of wiretap channels using Polar codes[J]. IEEE Transactions on Information Theory, 2011, 57(10): 6428-6443.

[19] WILDE M M, GUHA S. Polar codes for degradable quantum channels[J]. IEEE Transactions on Information Theory, 2011, 59(7): 4718-4729.

[20] SEIDL M, SCHENK A, STIERSTORFER C, et al. Multilevel polar-coded modulation[C]//IEEE International Symposium on Information Theory, July 7-12, 2013, Istanbul, Turkey. Piscataway: IEEE Press, 2013: 1302-1306.

[21] SEIDL M, SCHENK A, STIERSTORFER C, et al. Po-lar-coded modulation[J]. IEEE Transactions on Communications, 2013, 61(10): 4108-4119.

[22] CRONIE H S, KORADA S B. Lossless source coding with polar codes[C]//IEEE International Symposium on Information Theory, June 13-18, 2010, Austin, Texas, USA. Piscataway: IEEE Press, 2010: 904-908.

[23] ARIKAN E. Channel combining and splitting for cutoff rate improvement[J]. IEEE Transactions on Information Theory, 2006, 52(2): 628-639.

[24] 3GPP. Final minutes of 3GPP TSGRAN WG1 meet-ings# AH1_NR[R]. 2017.

[25] FORNEY G D. Codes on graphs: normal realizations[J]. IEEE Transactions on Information Theory, 2001, 47(2): 520-548.

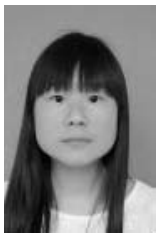
[26] GOELA N, KORADA S B, GASTPAR M. On LP decoding of polar codes[C]//IEEE Information Theory Workshop, October

16-20, 2010, Paraty, Brazil. Piscataway: IEEE Press, 2010: 1-5.
[27] TAL I, VARDY A. List decoding of polar codes[J]. IEEE Transactions on Information Theory, 2011, 61(5): 1-5.
[28] CHEN K, NIU K, LIN J R. List successive cancellation decoding of Polar codes[J]. Electronics Letters, 2012, 48(9): 500-501.
[29] NIU K, CHEN K. Stack decoding of polar codes[J]. Electronics Letters, 2012, 48(12): 695-697.
[30] CHEN K, NIU K, LIN J R. Improved successive cancellation decoding of Polar codes[J]. IEEE Transactions on Communications, 2013, 61(8): 3100-3107.
[31] 3GPP. Parity check bits for polar code: R1-1709996[S]. 2017.
[32] 3GPP. Polar codes construction for NR control channel: R1-1710048[S]. 2017.
[33] 3GPP. Polar code construction for NR control channel: R1-1710587[S]. 2017.
[34] 3GPP. CRC-based polar code construction: R1-1710490[S]. 2017.
[35] 3GPP. Polar code construction: R1-1710747[S]. 2017.
[36] CHUNG S Y, RICHARDSON T J, URBANKE R L. Analysis of sum-product decoding of low-density parity-check codes using a Gaussian approximation[J]. IEEE Transactions on Information Theory, 2001, 47(2): 657-670.
[37] 3GPP. Polar code design and rate matching: R1-167209[S]. 2016.
[38] 3GPP. Nested sequence for polar code: R1-1710000[S]. 2017.
[39] 3GPP. FRANK polar construction for NR control channel and performance comparison: R1-1709178[S]. 2017.
[40] 3GPP. Sequence construction of Polar codes for control channel: R1-1713468[S]. 2017.
[41] 3GPP. Evaluation of the sequence for polar codes: R1-1713469[S]. 2017.
[42] 3GPP. Design of combined-and-nested polar code sequences: R1-1710749[S]. 2017.
[43] 3GPP. Polar code size and rate-matching design for NR control channels: R1-1702735[S]. 2017.
[44] 3GPP. Polar rate-matching design and performance: R1-1713705[S]. 2017.
[45] 3GPP. Rate matching for polar code: R1-1711702[S]. 2017.
[46] 3GPP. Rate matching scheme for polar codes: R1-1713235[S]. 2017.
[47] 3GPP. Rate-matching scheme for control channel: R1-1711220[S]. 2017.
[48] 3GPP. Design of unified rate-matching for polar codes: R1-1710750[S]. 2017.
[49] 3GPP. Interleaver design for polar code: R1-1709999[S]. 2017.
[50] 3GPP. Polar code interleaver: R1-1712170[S]. 2017.
[51] 3GPP. Channel interleaver for polar codes: R1-1712649[S]. 2017.
[52] 3GPP. Interleaver design for NR polar codes: R1-1715835[S]. 2017.
[53] 3GPP. On interleaver pattern for downlink polar code construction: R1-1716487[S]. 2017.
[54] 3GPP. On downlink channel interleaver for polar codes: R1-1717996[S]. 2017.
[55] 3GPP. Downlink channel interleaver for polar codes: R1-1718678[S]. 2017.
[56] 3GPP. Performance comparison of interleaver design: R1-1711639[S]. 2017.
[57] 3GPP. Bit-interleaving for polar codes: R1-1716786[S]. 2017.

[作者简介]

谢德胜（1994−），男，重庆邮电大学移动通信重点实验室硕士生，主要研究方向为移动通信技术、软件定义网络和网络虚拟化。

柴蓉（1974−），女，重庆邮电大学移动通信重点实验室教授，主要研究方向为移动通信、软件定义网络、物联网、车联网体系架构、无线资源管理及移动性管理技术。



黄蕾蕾（1995−），女，重庆邮电大学移动通信重点实验室硕士生，主要研究方向为软件定义网络、无线资源管理和网络虚拟化。

陈前斌（1967−），男，重庆邮电大学副校长、教授、博士生导师，主要研究方向为个人通信、多媒体信息处理与传输和下一代移动通信网络等。

面向视频流的 MEC 缓存转码联合优化研究

李佳，谢人超，贾庆民，黄韬，刘韵洁，孙礼
（北京邮电大学，北京 100876）

摘　要：为应对未来移动网络所面临的巨大挑战，业界提出了自适应比特流（adaptive bit rate，ABR）技术和移动边缘计算（mobile edge computing，MEC），旨在为用户提供高体验质量、低时延、高带宽和多样化的服务。联合 ABR 和 MEC 来优化视频内容分发，对于提高网络性能和用户体验质量具有重要意义。其中，各项网络资源的联合优化是重要的研究课题。首先对 MEC 进行了概述，然后基于面向自适应流的 MEC 缓存转码联合优化问题，对业界已有工作进行了分析和对比，并对未来面临的挑战和研究难点进行了归纳和展望。

关键词：自适应视频流；MEC；缓存；转码
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018222

A survey on joint optimization of MEC caching and transcoding for video streaming

LI Jia, XIE Renchao, JIA Qingmin, HUANG Tao, LIU Yunjie, SUN Li
Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: To deal with the huge challenges in future mobile networks, the industry has proposed adaptive bit rate (ABR) technology and mobile edge computing (MEC), aiming to provide users with diverse services of high quality of experience, low latency and high bandwidth. Combining ABR and MEC to optimize the distribution of video content has been quite important for improving network performance and quality of experience. Especially, the joint optimization of network resources has arisen as an essential research topic. An overview of MEC was firstly given, and then the existing work in the industry on the joint optimization problem of MEC caching and transcoding oriented to adaptive streaming was analyzed and compared. Finally, the existing challenges in the future were summarized.

Key words: adaptive video streaming, MEC, caching, transcoding

1　引言

随着移动互联网技术的快速发展，移动网络数据流量呈现了爆发式的增长趋势，据 2017 年思科 VNI 报告，到 2021 年，全球移动数据流量将达到每月 49 EB（exabyte），其中移动视频流量将

收稿日期：2017−12−25；**修回日期**：2018−08−08
基金项目：中央高校基本科研业务费专项资金资助项目；国家自然科学基金资助项目（No.61501042）
Foundation Items：The Fundamental Research Funds for the Central Universities, The National Natural Science Foundation of China (No.61501042)

占全球移动数据流量的 78%[1]。爆发式增长的数据流量，尤其是视频流量给移动网络造成了巨大的挑战，要求移动网络具有提供更高数据传输速率和更低网络时延的能力，从而为用户提供更好的体验质量（quality of experience，QoE）。

为了应对上述网络挑战，一方面业界提出了在移动网络中采用自适应比特流（adaptive bit rate，ABR）传输技术进行视频分发[2]，即在移动网络中将视频文件编码为多种比特率版本，每个版本的视频文件都被切分成多个视频块（segment），并根据用户设备的能力、网络连接状况和特定的请求，为用户动态选择传送的视频块比特率版本，从而减少视频播放卡顿和重新缓冲率，提升用户的 QoE 体验[3]。另一方面，业界也提出了移动边缘计算（mobile edge computing，MEC），旨在移动网络边缘向内容提供商和应用开发者提供云计算能力和 IT 服务环境，从而为终端用户提供超低时延和高带宽的服务[4]。自 ABR 与 MEC 技术提出以后，通过利用 MEC 实现自适应视频流的缓存、转码与自适应分发受到了业界的广泛关注[5-6]，并表现出了显著的优势，主要体现在 MEC 在网络边缘提供了存储能力和计算能力，可实现视频内容的就近缓存，并根据动态变化的网络状况在网络边缘进行视频转码，以实现视频的自适应分发，从而降低视频内容的传输时延并节省网络带宽，同时提升用户的 QoE 体验。

目前已有大量工作将 MEC 与 ABR 技术联合考虑，并从不同的部署场景出发，针对视频流的缓存转码等问题进行了研究[7-18]，其中不仅包括单 MEC 服务器场景中的资源协同问题，还对分布式多 MEC 之间的资源协作问题进行了讨论，并针对不同的场景和需求，对缓存策略和转码策略进行了设计，从而优化包括系统成本、网络负载和视频容量在内的各项网络指标。另外，由于新型网络架构在增强视频业务质量方面展现的高效性能，部分研究工作还面向视频流分发业务提出了基于 SDN 的 MEC 架构，从而优化网络架构和视频分发机制。

尽管目前对 MEC 中缓存转码问题的研究已取得了一定的进展，但尚未有系统性的工作对基于 MEC 的自适应视频流分发研究的核心关键技术问题、业界已取得的进展以及未来仍面临的挑战等问题做全面的分析与总结。因此，本文将针对面向视频流的缓存转码联合优化问题，对目前已有工作进行总结和对比，并对目前仍需解决的问题进行归纳和展望。

2　自适应视频流场景下的 MEC 部署架构

欧洲电信标准化协会（European Telecommunications Standards Institute，ETSI）于 2014 年首次提出了移动边缘计算，并且给出了其“在移动网络边缘提供 IT 服务环境和云计算能力”的定义。MEC 将原本位于云数据中心的服务和功能下沉到网络边缘，提供超低时延、超高带宽的网络环境和实时网络分析能力。由于在回传网和核心网中发挥的低时延保证和容量增强的重要作用，MEC 技术受到了广泛关注，被公认为是 5G 的主要关键技术之一[19]。

ETSI MEC 基于虚拟化平台部署，目标是为第三方应用提供标准的体系架构和符合行业标准的 API，使得各种应用可以运行在网络边缘，同时基于 NFV（network function virtualization，网络功能虚拟化）的虚拟化基础设施有利于网络运营商的重复利用。MEC 服务器可以部署在网络边缘的不同位置，如 LTE 宏基站（eNode B）、3G 无线网络控制器（radio network controller，RNC）、汇聚点或核心网侧等[20]。

MEC 服务器通常具有较高的计算能力，适合于分析处理大量数据，可以为 AR 和 VR 等计算密集型业务提供高效的计算资源。另外，由于 MEC 在地理上十分接近用户或信息源，可以显著降低业务响应的时延，并减小回传网和核心网发

生网络拥塞的可能性。最后，位于接入网侧的MEC 能够实时感知网络数据，包括无线链路状况和用户行为信息及位置信息等，从而进行链路感知自适应，极大地改善用户的服务质量体验[21-22]。

与传统的网络架构相比，MEC 具有许多显著的优势，可以有效解决传统网络模式中高时延和低效率等问题。在面向视频流业务时，MEC 的存储计算资源以及网络感知能力可以有效地支持ABR 技术。一方面，MEC 的分布式边缘存储资源可以对视频内容进行缓存，并且对视频请求进行本地卸载，从而既可以缩短用户到视频内容的距离，降低传输时延，也可以减少视频内容的冗余传输，节省网络带宽并提高能量效率，缓解核心网络压力，另一方面，在 ABR 中，用户请求的视频比特率可基于网络条件、设备功能和用户偏好自适应调整[23]，因此往往需要对视频块进行多个比特率版本的缓存，而对一个视频块缓存多种比特率版本会造成较大的缓存成本，因此可以选择在 MEC 中缓存一部分较高比特率视频，对于缓存不命中请求，利用 MEC 的计算能力使用视频转码技术将缓存的视频版本转码为请求的比特率版本，视频转码可以提升缓存资源的利用率，在基于 ABR 的视频分发系统中具有重要的作用，此外，MEC 还可以利用其网络感知能力，对用户和网络信息进行分析，有效支持 ABR 技术。在自适应视频流业务中利用 MEC 缓存流行视频块，并在不同比特率版本之间进行转码，实现在网络边缘对用户请求进行响应，已经被认为是解决 ABR 中视频内容分发的一个重要趋势。

MEC 的缓存和转码能力可以有效支持 ABR 视频业务，从而缓解源服务器压力并提升视频观看质量，但在 MEC 资源容量有限的情况下，缓存转码资源的滥用会造成 MEC 负载过重，降低资源效率，影响用户体验，因此需要对不同视频业务的需求和特点进行分析，在各项资源之间进行权衡和分配，使得视频缓存转码的效率最优。除此之外，由于 MEC 分布式部署的特点，可以合理利用邻近 MEC 服务器的缓存转码资源，协作处理视频请求，优化整个网络的视频分发质量和资源效率。因此，研究 MEC 中面向视频流的计算、缓存与网络资源的联合优化问题，均衡地对缓存、计算和带宽资源进行合理编排，对优化系统整体性能具有重要意义。

图 1 为一个自适应视频流场景下的 MEC 部署架构，MEC 服务器的位置部署在基站之后，其中部署了缓存和计算资源，可以对视频文件进行缓存和转码，当用户的请求到达 MEC 时，若缓存命中或转码命中，就可以在本地对请求进行响应，如图 1 中的(a)路线和(b)路线，而不需要经过图 1 中(c)路线所示的回传网和核心网的传输，从源服务器获取视频文件。这样一来，既显著降低了用户请求响应的时延，保证用户体验，同时还可以避免网络拥塞，节省回传网和核心网的资源。

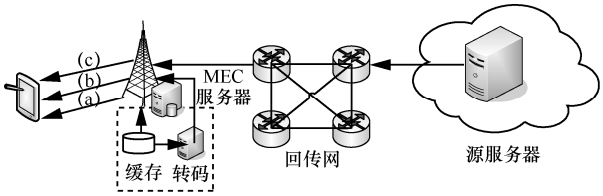


图 1　自适应视频流场景下的 MEC 部署架构

3　面向视频流的缓存转码资源联合优化

利用 MEC 实现 ABR 技术，可以有效解决视频内容的缓存转码和分发问题，其中缓存转码资源的联合优化已经被认为是一个重要的研究课题。在传统的视频流分发系统中，缓存和转码都在云端完成，由于传统方案中较高的传输时延和回传压力，MEC 逐渐得到了广泛的应用，研究结果表明部署 MEC 可以有效提升系统性能和用户体验。

3.1　在云端的缓存转码联合优化

媒体云作为当前一种有效实现自适应流分发的框架，在基于云的体系架构中，动态按需编排

虚拟化存储和计算资源，为显著降低系统成本提供了可能性，图 2 为在媒体云中应用自适应视频流业务的缓存转码示意图，媒体云的流媒体引擎从媒体库中获取视频内容，并进行视频内容的缓存和转码，进而实现自适应视频流向用户的分发。Jin 等人[7]研究了媒体云中缓存和转码的最优化策略问题，旨在动态调度各种资源，从而最小化某个节点响应单个请求的总运营成本。具体来说，在缓存策略方面，该方案考虑了视频的流行度，即对流行度最高的若干视频内容缓存全部比特率版本，对流行度次高的一部分视频内容缓存最高比特率版本。当在节点缓存命中时，直接利用缓存响应请求；当转码命中时，比较转码成本和带宽成本，选择成本较小的方式提供服务；当本地缓存不可用时，则从源服务器获取内容。参考文献[7]将缓存、转码和带宽成本之间的权衡问题建模成一个凸优化问题，然后采用两步分析法，分别对缓存资源分配和转码配置策略进行优化，最后得到最优的节点缓存空间大小和所有版本都缓存的视频数量。

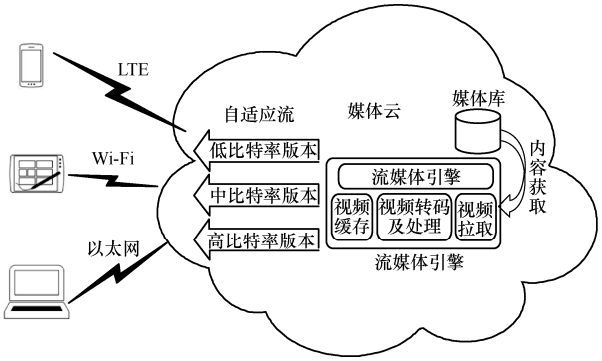


图 2　媒体云中面向自适应视频流的缓存与转码

面向视频流的资源联合优化，不仅可以从网络资源配置的角度出发，还可以从用户请求内容的角度进行视频块的调度决策。Gao 等人[8]根据对用户视频行为的分析，研究了媒体云中视频内容的管理问题，提出了一种成本高效的部分转码方案。具体来说，每个视频被分为一组片段，其中一部分片段被预处理并存储在缓存中，而其他片段则在播放过程中被快速转码。缓存策略方面，服务器中存储着源视频文件和一部分转码版本，源文件可以转码为任何比特率的版本，另外基于用户实时变化的比特率需求和部分转码的方案，采取动态缓存策略，从而在每个时隙根据用户请求信息来决定是否缓存。当用户请求在节点缓存命中时，用户可以直接获取缓存中的视频内容，进而产生存储成本，缓存不命中时，则进行在线转码，进而产生计算成本。该方案考虑了与存储和计算相关的总成本，建立了一个约束随机优化问题，优化目标为最小化某个服务器的长期总成本，决定某个视频段应缓存还是在线转码，然后利用李雅普诺夫优化框架和拉格朗日松弛法，设计了在线算法进行求解。

通过对用户视频行为的分析可以看出，超过 60%的视频只有前 20%的内容被用户观看，视频内部存在着不同的流行度，从这一角度出发可以进一步优化网络资源调度。Zhao 等人[9]研究了云网络中缓存和转码的均衡问题，基于对视频的分段，考虑了视频段之间的不同流行度以及云中存储和计算价格的大小，对系统中所有视频的放置问题进行了研究，从而决定对每个视频段的哪些版本进行缓存或者转码，将总的系统运营成本最小化。具体来说，首先根据视频内部流行度将视频文件分为多个视频段，通过一个转码权重图描述不同比特率版本之间的转码关系，从而计算出各版本之间的转码成本。基于以上的考虑，提出了一个存储和转码的权衡策略，对流行视频段存储多个或者所有版本，对不流行的视频段存储最高比特率版本，并针对用户请求进行转码。为了避免转码启动时延对视频播放效果的影响，可以在播放视频段时，提前对下一个视频段进行转码。该方案将该权衡问题描述为一个优化问题，优化目标为最小化与存储和转码相关的总成本，并利用启发式分治算法求解，进行视频段某个版本的

缓存和转码决策。

3.2 在 MEC 中的缓存转码联合优化

在云端进行缓存和转码，会带来较高的传输时延，增加回传网和核心网的带宽压力，而 MEC 逐步展现出应对视频业务挑战的优势，利用 MEC 的存储和计算能力，可以有效提高网络的整体性能，并显著改善用户体验。其中，MEC 中缓存和转码的联合优化问题已成为一个重要的研究课题。

图 3 所示为在分布式 MEC 架构中实现自适应视频流业务的示意，其中，每个 MEC 服务器都可以实现视频的缓存和转码，分布式部署的 MEC 服务器之间可以实现协作式的缓存和转码，从而实现资源的高效利用，进一步提升视频分发效率和用户体验。

为解决无线网络中自适应比特流的缓存挑战，Pedersen 等人在参考文献[10-11]中研究了自适应比特流场景中无线缓存和处理的联合优化问题。该方案首先将视频文件分为多个视频块，每个视频块可以按不同的比特率请求。针对 RAN 的缓存挑战，提出在 RAN 部署有限的计算资源，从而可以进行视频块之间的转码，缓解存储压力。基于以上考虑，提出了基于 ABR 感知的主动/被动的联合转码和缓存资源的策略，具体来说，对于视频请求有 3 种内容获取方式，从缓存处直接获取对应版本，对缓存的高比特率版本进行转码，或者通过回传网络从 CDN 处获取。当用户请求在节点缓存命中时，直接由缓存进行响应，缓存不命中时，可以根据给定可用的缓存容量、处理能力和回传带宽，通过转码资源和回传资源分配算法进行转码和回传决策，当采用回传方式时，采用缓存策略对获取的内容进行缓存。参考文献[10-11]中分别采用了两种缓存方式，分别是 LRU（least recently used）缓存策略和 P-UPP（proactive user preference profile）缓存策略。该方案制定了一个优化问题，优化目标为最大化无线网络的视频容量，即服务的并发视频请求数量，并采用启发式算法进行了求解，从而对某个视频块的获取方式进行调度决策。

Wang 等人[12-13]提出了一个在线转码和地域分布式交付的联合策略，系统架构中包括多个 CDN 区域，每个区域中包含后端服务器和对等服务器，转码任务在后端服务器中完成。该方案考虑了用户的 CDN 区域偏好、区域的转码版本偏好以及视频请求的用户偏好。首先根据用户的 CDN 区域偏好，即考虑服务器到用户的带宽大小对用户进行重定向，选择提供服务的 CDN 区域，该区域中的对等服务器以循环方式提供服务。另外，根据内容的用户偏好和区域的转码版本偏好来安排转码任务，并选择空闲的 CDN 计算资源进行转

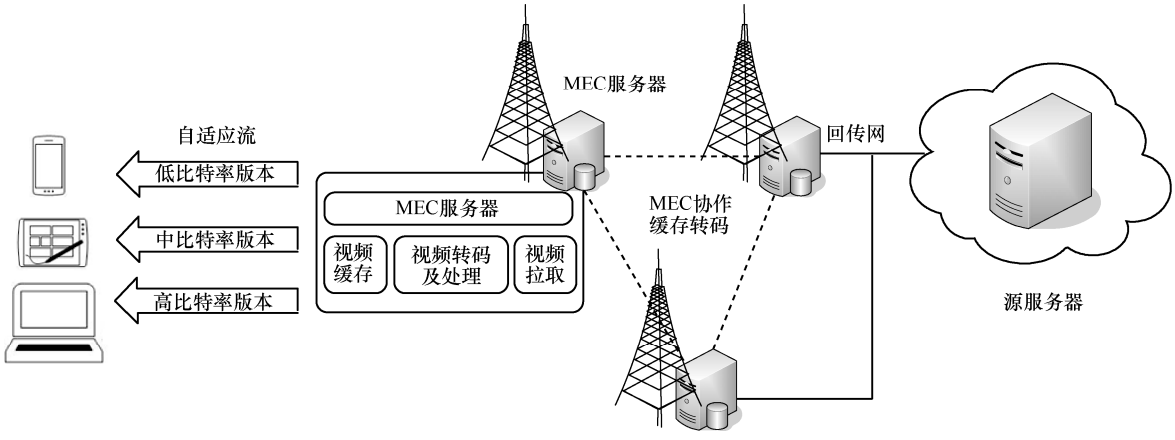


图 3 分布式 MEC 架构中面向自适应视频流的协作缓存与转码

码和交付，减少跨区域的复制成本，特别地，根据按需设计的策略，视频段被转码为一组预定义的版本，如果转码不及时，可以选择最接近的比特率版本进行转发。最后，该方案对优化问题进行了建模，优化目标为最小化计算成本和复制成本，并采用启发式和分布式算法进行了求解。

在视频业务中，当视频比特率和传输条件不匹配时，会引发网络的拥塞和较高的时延，造成视频播放的卡顿，严重影响用户的观看体验，因此，在资源联合优化的过程中，不仅要从运营商的角度出发考虑成本代价，还要从用户的角度出发保障 QoE。众包直播游戏视频流（crowdsourced live game video streaming，CLGVS）是一种新兴的互联网业务，可以使众多异构终端随时随地观看游戏玩家播放的视频，Zheng 等人[14]研究了 CLGVS 业务中的在线转码和交付问题，从 CLGVS 服务提供商的角度，通过联合优化动态转码决策、比特率配置和数据中心选择，减少运营成本并保证用户的服务质量。该方案考虑了两种转码策略，分别为门限转码策略和全部转码策略。该问题被建模为一个约束随机优化问题，优化目标包括两部分，即最小化与计算和带宽相关的总运营成本，最大化与时延和比特率相关的用户 QoE。然后利用李雅普诺夫优化框架，设计了一个在线算法 OCTAD（online cloud transcoding and distribution）进行求解，算法包括 3 个主要部分：动态直播转码决策、自适应比特流配置和智能数据中心选择，从而动态地为每个游戏玩家执行比特率配置，为每个观看者进行转码决策和数据中心选择。同时，为了扩展算法的适用范围，还在设计在线算法时考虑了游戏的类型。

现有的缓存转码优化方案主要基于单服务器独立进行缓存和转码任务决策的场景，没有考虑服务器之间的协作，为了研究在多 MEC 服务器场景下联合优化整体运营成本的问题，Tran 等人[15-16]提出了一种移动边缘计算网络中多比特率视频流的协作缓存和处理策略，称为 CoPro-CoCache。在该方案中，获取视频内容的方法可能有：从本地服务器的缓存中获取，在本地服务器中转码获得，从协作服务器的缓存中获取，在协作服务器中进行转码并传回，从协作服务器传回后在本地进行转码。具体来说，缓存策略方面，本方案不需要内容流行度的先验信息，采用 LRU 缓存策略，将每个小区最流行的视频缓存在对应的基站缓存服务器上，直到缓存存储空间已满，当用户的视频请求需要对缓存中的比特率版本进行转码时，将转码任务分配给负载最小的 MEC 服务器，从而均衡网络负载，这个服务器可以是存储原始版本的 MEC 服务器，即数据提供节点，也可以是交付节点。参考文献[16]将该协作缓存和处理问题建模为一个整数线性规划问题，该问题受存储空间和处理能力的约束，在给定可用资源后，优化目标为对单个视频请求协作制定缓存放置策略和视频调度策略，从而最小化回传网络成本。最后，针对该 NP 问题提出了一个新型的在线算法 JCCP（joint collaborative caching and processing）来进行求解。

Xu 等人[17]提出了 MEC 增强的自适应比特率（MEC-ABR）视频传输方案，联合进行缓存和无线资源的分配。在该方案中，MEC 服务器作为控制组件来执行缓存策略并灵活调整视频版本。具体来说，该方案首先考虑了 BS 的流量负载，从而进行 MEC 服务器的存储资源分配，以缓存各 BS 服务范围内的流行视频，并将该存储资源分配问题模拟为 Stackelberg 博弈进行了求解。缓存策略方面，不仅考虑了视频的流行度，还考虑了 RAN 侧的无线信道质量，缓存策略和视频交付可以被灵活地调整，以匹配不同的无线信道。另一方面，该方案将联合缓存和无线资源的分配问题建模为匹配问题，BS 和用户分别根据视频的流行度和无线信道条件维护偏好列表，利用 MEC 的存储和计算能力进行优化，提出了 JCRA（joint cache and

radio resource allocation）算法来解决这个问题，并考虑了视频比特率版本的动态调整。

软件定义移动网络（software-defined mobile network，SDMN）、网内缓存和 MEC 作为下一代移动网络的重要技术，对增强视频业务质量具有重要意义，Liang 等人[18]研究了一个 MEC-SDMN 中的视频速率适应问题，联合考虑视频速率自适应、带宽配置和 MEC 中的计算资源调度，设计了一个高效机制。研究目的是在考虑网络资源和视频缓存分布的情况下，为每个用户找到最佳的视频质量水平。在该方案中，SDN 控制器执行流量管理，通过最大限度地增大视频的整体平均增益，提高整个网络的效用，以帮助用户自适应地选择最佳的视频质量水平。为了最大化 HetNet 的平均视频质量，该方案制定了一个优化问题，并采用双分解方法，将视频数据速率、计算资源和流量管理（带宽配置和路径选择）3 部分问题解耦，独立求解各个变量。

3.3 方案对比

本章前两部分主要从云端和 MEC 两个场景出发，介绍了目前已有工作中面向视频流的缓存、计算和带宽资源的联合优化方案。针对以上方案策略和优化方法等的不同，下面从缓存策略、转码策略、建模方法和算法等具体方面对已有方案进行了对比，见表 1。

从表 1 可以看出，已有工作主要面向的是单服务器架构，对协作缓存转码问题考虑得较少。缓存方面主要采用一部分全部缓存和一部分缓存最高版本的策略，在考虑视频流行度的情况下，可以缓存流行视频的全部版本，而对不流行视频缓存最高比特率版本。转码策略包括以下几种策略：缓存不命中直接转码，与带宽资源做均衡进行转码决策，协作转码中考虑负载和成本进行转码决策等。另外，建立优化问题时，优化目标主要为最小化系统成本，除此之外，还包括对视频容量和 QoE 等方面的优化。解决问题的算法主要包括李雅普诺夫优化理论、分析法、博弈论、启发式算法和在线算法等。

4 面向视频流的 MEC 资源优化问题与挑战

在网络边缘部署 MEC 来应对视频流业务，可以有效降低传输时延并节省网络资源。虽然目前已有大量工作对 MEC 面向 ABR 的缓存、计算和带宽资源的分配和调度问题进行了研究，但如何综合考虑网络各方面因素，均衡各项资源，从而使系统整体性能最优，有多方面的挑战和研究难点，本文总结了以下 5 个方面。

4.1 缓存转码带宽资源优化与 QoE 优化的均衡问题

QoE 是一种以用户认可程度为标准的服务评价方法，直接反映了用户在一定客观环境中对适用的服务或业务的整体认可程度[24]。自适应流媒体的发展是推动探索增强 QoE 的有效方法的关键驱动力，从而通过对用户提供差异化服务来保障用户体验[25]。MEC 中缓存转码带宽资源的优化与 QoE 优化的均衡问题是一个非常有意义的研究方向，从视频内容提供商的角度出发，对于系统的优化一般要考虑两个维度：一方面要降低缓存、计算和网络的运营成本，另一方面又要保证终端用户的 QoE。因此，如何权衡 MEC 缓存转码带宽资源的租赁成本与终端用户的 QoE 保证，是今后研究的一个重要方向。

4.2 缓存转码带宽资源的能量效率优化问题

MEC 的部署将原本位于云端的存储和计算资源下沉到网络边缘，一方面使得网络边缘可以对用户请求进行响应，一方面可以减少回传资源的浪费。在 MEC 的网络优化方面，能量效率问题是关注的重点问题之一。在 MEC 的部署场景中，内容的缓存、MEC 的计算以及 MEC 之间、MEC 与用户之间的通信都会产生大量的能耗，从而带来极大的能耗成本。因此，建立能量高效的资源优化机制，对缓存、计算和通信资源进行有效的调度，对于减少系统能耗、提高系统性能有着重

表 1　缓存转码联合优化方案对比

	参考文献	架构	缓存策略	转码策略	建模	解决问题	算法	偏好考虑
云端	[7]	单服务器	缓存最流行的视频的所有版本，缓存不流行视频的最高版本	对缓存不命中，有最高版本缓存，且转码成本低于带宽成本的请求进行转码	凸优化问题 优化目标：最小化与缓存、计算和带宽相关的总运营成本	资源配置：服务器缓存资源分配和转码策略	两步分析法	视频流行度
	[8]	单服务器	缓存最高版本的源视频文件以及一部分转码版本	部分转码	约束随机优化问题 优化目标：最小化与存储和计算相关的长期总成本	某视频段的缓存和转码策略	利用李雅普诺夫优化框架和拉格朗日松弛法设计的在线算法	无
	[9]	单服务器	流行视频块缓存多个版本，不流行视频块缓存最高比特率版本	缓存不命中时进行转码，还可以提前对下一个视频段进行转码	最小化存储和转码相关的总成本	所有视频段的缓存和转码策略	启发式分治算法	视频段流行度
MEC	[10] [11]	单服务器	LRU 缓存策略 P-UPP 缓存策略	通过转码资源和回传资源分配算法进行转码决策	多背包问题 优化目标：最大化网络视频容量和 QoE	某视频块的调度策略	启发式算法	无
	[12] [13]	多区域多服务器协作	缓存视频块的最高比特率版本	协作转码	优化目标：减少计算资源消耗，最小化复制成本	转码资源的调度	启发式和分布式算法	请求用户偏好、区域用户偏好和转码版本区域偏好
	[14]	多服务器无协作	缓存最高比特率版本的视频	门限转码策略和全部转码策略	约束随机优化问题 优化目标：最小化与计算和带宽相关的总运营成本，最大化与时延等相关的用户 QoE	为每个游戏玩家执行比特率配置，为每个观看者进行转码决策和数据中心选择	利用李雅普诺夫优化框架设计的在线算法 OCTAD	无
	[15] [16]	多服务器协作	缓存最流行的视频块，采用 LRU 缓存策略，协作缓存	协作转码，选择负载最小的服务器进行转码	整数线性规划问题 优化目标：最小化回传网络成本，受缓存和处理能力约束	某视频块的缓存放置策略和调度策略	JCCP 在线算法	无
	[17]	单服务器	缓存各小区流行视频	按需转码	MEC 缓存资源分配：Stackelberg 博弈 联合缓存和无线资源分配：匹配问题	资源分配问题：缓存资源和无线资源	Stackelberg 博弈论；JCRA 在线算法	视频流行度
	[18]	多服务器	缓存一部分视频的最高版本	根据计算容量进行转码决策	优化目标：最大化平均视频质量水平	为每个用户寻找最佳视频质量	双分解方法	无

要意义。在面向视频流的 MEC 缓存转码带宽资源的联合优化中，主要关注缓存能耗、转码能耗和传输能耗，如何联合考虑 MEC 的计算、转码和传输，优化提供视频流业务的能量效率是今后研究的一个重点。

4.3　MEC 中基于深度增强学习的缓存和转码

深度增强学习是将深度学习和增强学习结合起来，从而实现端对端学习的一种全新的算法，是通用的人工智能框架，目前已经成为网络优化的重要方法和工具[26]。结合深度增强学习的方法，对自适应视频流内容进行缓存是一个重要研究方向。在基于 ABR 视频流缓存系统中，每个视频块都有多个比特率版本，考虑到边缘网络缓存系统的容量限制，缓存所有比特率的视频块会造成缓

存资源利用率的降低和网络成本的增加。通过部署在网络无线接入侧的 MEC，可以实时对网络信息进行感知，包括网络链路状况和用户行为等数据[22]，利用深度增强学习的方法对这些信息进行分析和学习，可以预测视频内容的流行度以及用户对响应视频块比特率版本的请求状况，提前进行资源分配和调度，对相应视频内容和比特率版本进行缓存，从而提高缓存命中率和缓存资源的利用率。

4.4 分布式的多 MEC 协作问题

MEC 在边缘网络中的部署，通常采用分布式的方式，因此 MEC 带来的缓存和计算资源也分布式的位于网络的不同位置。单个 MEC 的存储空间和计算能力都是有限的，过多的缓存和计算任务会给 MEC 服务器造成过载，而回传到云数据中心又会产生较高的回传成本，因此基于 MEC 分布式的部署方式，相邻的 MEC 服务器之间可以协作进行缓存和计算，当前 MEC 服务器没有相应缓存内容或计算资源紧张时，可以调用其他空闲 MEC 服务器，此类分布式协作的方式可以有效减少网络运营成本，提高网络性能[27]。因此，不同 MEC 节点之间如何协作共享资源（主要包括计算和缓存资源）成为一个重要的研究问题。例如，当用户请求的目标视频内容在本地 MEC 服务器没有缓存时，如何在其他缓存有相应内容的 MEC 节点中选择一个最优的节点；当本地 MEC 服务器的计算负荷过载时，如何将本地的计算任务卸载至其他的 MEC 节点，这都需要 MEC 节点之间的协作。因此研究基于分布式的多 MEC 协作的资源共享机制，以提高资源的利用率和用户的体验也是今后进行资源联合优化的一个重要方向。

4.5 基于 NFV、SDN 和网络切片等新型技术的资源分配问题

在自适应视频流场景中，不同的网络环境和用户能力可以动态适配视频流的比特率版本，而不同类型的视频业务和不同等级的用户对 QoE 保障有着差异化的需求，另一方面，由于单 MEC 服务器的计算和存储资源有限，分布式多 MEC 场景下的资源协同也面临着很大的需求和挑战[28]。如何针对不同业务场景，利用新型网络技术实现资源的高效管理和分配是 MEC 中面向视频流业务的重要研究课题。网络切片技术可以针对不同应用场景，将物理网络切割成多个虚拟网络，从而应对不同场景中对传输时延、移动性、可靠性、安全性以及计费方式的差异性，利用边缘计算的计算、存储和通信能力，构建业务所在无线接入网络内的接入网切片，可以实现业务的本地处理，缓解核心网压力，减少传输时延，改善业务性能[29]。除此之外，未来的 5G 网络还提出了如下演进目标：基于 SDN/NFV 进行虚拟化，进行扁平化扩展与增强，其中 NFV 和 SDN 是实现网络切片的基础，NFV 提供了按需分配的可配置资源共享池，可以极大地方便资源的统一管理，同时 SDN 实现了集中式的控制平面，并通过为用户提供的编程接口，使用户可以根据上层业务和应用个性化地定制网络资源来满足其特有的需求[30]。针对不同的业务场景，进行有效的网络业务切片和划分，并通过 SDN 的全局管控和对 NFV 虚拟资源的合理调配，对优化整体资源效率和网络性能具有重要研究意义。

5 结束语

本文从 MEC 和 ABR 的背景和概述出发，对目前面向视频流的缓存转码资源联合优化方案进行了介绍和分析，并主要从缓存策略、转码策略和优化方式等方面对已有方案进行了对比。在对以上方案分析对比的基础上，研究了面向视频流的 MEC 资源优化问题目前面临的挑战和研究难点，如与时延优化的均衡问题、能量优化问题和用户行为分析以及分布式 MEC 协作的问题等，在网络整体优化方面具有重要意义。

参考文献：

[1] Cisco Mobile VNI. Cisco visual networking index: global mobile data traffic forecast update, 2016–2021 white paper[R]. 2017.

[2] STOCKHAMMER T. Dynamic adaptive streaming over HTTP: standards and design principles[C]//The Second Annual ACM Conference on Multimedia Systems, Feb 23-25, 2011, San Jose, CA, USA. New York: ACM Press, 2011: 133-144.

[3] YIN X Q, JINDAL A, SEKAR V, et al. A control-theoretic approach for dynamic adaptive video streaming over HTTP[C]//The 2015 ACM Conference on Special Interest Group on Data Communication Pages, August 17-21, 2015, London, UK. New York: ACM Press, 2015: 325-338.

[4] ETSI. Mobile edge computing—a key technology towards 5G[R]. 2015.

[5] LI Y, FRANGOUDIS P A, HADJADJ-AOUL Y, et al. A mobile edge computing-based architecture for improved adaptive HTTP video delivery[C]//2016 IEEE Conference on Standards for Communications and Networking (CSCN), Oct 29-31, 2016, Paris, France. Piscataway: IEEE Press, 2016: 1-6.

[6] WANG C C, LIN Z N, YANG S R, et al. Mobile edge computing-enabled channel-aware video streaming for 4G LTE[C]//Wireless Communications and Mobile Computing Conference, June 26-30, 2017, Valencia, Spain. Piscataway: IEEE Press, 2017: 564-569.

[7] JIN Y, WEN Y, WESTPHAL C. Optimal transcoding and caching for adaptive streaming in media cloud: an analytical approach[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(12): 1914-1925.

[8] GAO G, ZHANG W, WEN Y, et al. Towards cost-efficient video transcoding in media cloud: insights learned from user viewing patterns[J]. IEEE Transactions on Multimedia, 2015, 17(8): 1286-1296.

[9] ZHAO H, ZHENG Q, ZHANG W, et al. A segment-based storage and transcoding trade-off strategy for multi-version VoD systems in the cloud[J]. IEEE Transactions on Multimedia, 2017, 19(1): 149-159.

[10] AHLEHAGH H, DEY S. Adaptive bit rate capable video caching and scheduling[C]//IEEE WCNC'13, April 7-10, 2013, Shanghai, China. Piscataway: IEEE Press, 2013: 1357-1362.

[11] PEDERSEN H A, DEY S. Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing[J]. IEEE/ACM Transactions on Networking, 2016, 24(2): 996-1010.

[12] WANG Z, SUN L, WU C, et al. Joint online transcoding and geo-distributed delivery for dynamic adaptive streaming[C]//IEEE INFOCOM'14, April 29-May 1, 2014, Toronto, Canada. Piscataway: IEEE Press, 2014: 91-99.

[13] WANG Z, SUN L, WU C, et al. A joint online transcoding and delivery approach for dynamic adaptive streaming[J]. IEEE Transactions on Multimedia, 2015, 17(6): 867-879.

[14] ZHENG Y, WU D, KE Y, et al. Online cloud transcoding and distribution for crowdsourced live game video streaming[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(8): 1777-1789.

[15] TRAN T X, HAJISAMI A, PANDEY P, et al. Collaborative mobile edge computing in 5G networks: new paradigms, scenarios, and challenges[J]. IEEE Communications Magazine, 2017, 55(4): 54-61.

[16] TRAN T X, PANDEY P, HAJISAMI A, et al. Collaborative multi-bitrate video caching and processing in mobile-edge computing networks[C]//WONS'13, March 18-20, 2017, Banff, Canada. [S.l.:s.n.], 2017: 165-172.

[17] XU X, LIU J, TAO X. Mobile edge computing enhanced adaptive bitrate video delivery with joint cache and radio resource allocation[J]. IEEE Access, 2017(5): 16406-16415.

[18] LIANG C, HU S. Dynamic video streaming in caching-enabled wireless mobile networks[J]. arXiv: 1706.09536, 2017.

[19] 王胡成, 徐晖, 程志密,等. 5G 网络技术研究现状和发展趋势[J]. 电信科学, 2015, 31(9): 149-155.
WANG H C, XU H, CHENG Z M, et al. Current research and development trend of 5G network technologies [J]. Telecommunications Science, 2015, 31(9): 149-155.

[20] 李子姝, 谢人超, 孙礼, 等. 移动边缘计算综述[J]. 电信科学, 2018, 34(1): 87-101.
LI Z S, XIE R C, SUN L, et al. A survey of mobile edge computing[J]. Telecommunications Science, 2018, 34(1): 87-101.

[21] TALEB T, SAMDANIS K, MADA B, et al. On multi-access edge computing: a survey of the emerging 5G network edge cloud architecture and orchestration[J]. IEEE Communication Surveys & Tutorials, 2017, 19(3): 1657-1681.

[22] WANG S, ZHANG X, ZHANG Y, et al. A survey on mobile edge networks: convergence of computing, caching and communications[J]. IEEE Access, 2017, 5(3): 6757-6779.

[23] TIMMERER C, GRIWODZ C. Dynamic adaptive streaming over HTTP: from content creation to consumption[C]//Proc. of ACM MM'12, Oct 29-Nov 2, 2012, Nara, Japan. New York: ACM Press, 2012: 1533-1534.

[24] 赵希鹏，张欣，杨大成，等. 基于 QoE 的无线网络资源调度优化研究[J]. 移动通信, 2014(22): 8-13.
ZHAO X P, ZHANG X, YANG D C, et al. Research on optimization of wireless network resource scheduling base on QoE[J]. Mobile Communications, 2014(22): 8-13.

[25] LI C, TONI L, ZOU J, XIONG H, et al. QoE-driven mobile edge caching placement for adaptive video streaming[J]. IEEE Transactions on Multimedia, 2017(9): 1.

[26] HE T Y, ZHAO N, YIN H. Integrated networking, caching and computing for connected vehicles: a deep reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2017, 99(10): 1.

[27] GHARAIBEH A, KHREISHAH A, JI B, et al. A provably efficient online collaborative caching algorithm for multicell-coordinated systems[J]. IEEE Transactions on Mobile Computing, 2016, 15(8): 1863-1876.

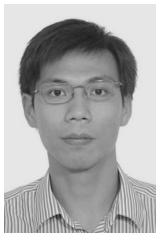
[28] WANG W, CAO J, ZHANG W. Edge computing: vision and challenges[J]. IEEE Internet of Things Journal, 2016, 3(5): 637-646.

[29] 项弘禹, 肖扬文, 张贤, 等. 5G 边缘计算和网络切片技术[J].

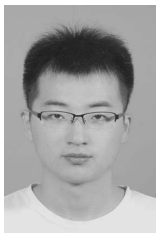
电信科学, 2017, 33(6): 54-63.
XIANG H Y, XIAO Y W, ZHANG X, et al. Edge computing and network slicing technology in 5G[J]. Telecommunications Science, 2017, 33(6): 54-63.
[30] GPPP E B. QoE-oriented mobile edge service management leveraging SDN and NFV[J]. Mobile Information Systems, 2017(1).

[作者简介]

李佳（1994−），女，北京邮电大学未来网络理论与应用实验室硕士生，主要研究方向为 5G 网络、移动边缘计算等。



谢人超（1984−），男，北京邮电大学未来网络理论与应用实验室副教授、硕士生导师，主要研究方向为信息中心网络、移动网络内容分发技术和移动边缘计算等。



贾庆民（1990−），男，北京邮电大学未来网络理论与应用实验室博士生，主要研究方向为新型网络体系架构、内容分发和移动边缘计算等。

黄韬（1990−），男，北京邮电大学未来网络理论与应用实验室教授、博士生导师，主要研究方向为新型网络体系架构、内容分发网络软件定义网络等。

刘韵洁（1943−），男，中国工程院院士，北京邮电大学教授、博士生导师，主要研究方向为未来网络体系架构。

孙礼（1959−），男，北京邮电大学未来网络理论与应用实验室副教授、硕士生导师，主要研究方向为宽带通信网络、无线接入技术、通信网络交换技术等。

面向视频流的 MEC 缓存转码联合优化研究

李佳，谢人超，贾庆民，黄韬，刘韵洁，孙礼
（北京邮电大学，北京 100876）

摘　要：为应对未来移动网络所面临的巨大挑战，业界提出了自适应比特流（adaptive bit rate，ABR）技术和移动边缘计算（mobile edge computing，MEC），旨在为用户提供高体验质量、低时延、高带宽和多样化的服务。联合 ABR 和 MEC 来优化视频内容分发，对于提高网络性能和用户体验质量具有重要意义。其中，各项网络资源的联合优化是重要的研究课题。首先对 MEC 进行了概述，然后基于面向自适应流的 MEC 缓存转码联合优化问题，对业界已有工作进行了分析和对比，并对未来面临的挑战和研究难点进行了归纳和展望。

关键词：自适应视频流；MEC；缓存；转码
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018222

A survey on joint optimization of MEC caching and transcoding for video streaming

LI Jia, XIE Renchao, JIA Qingmin, HUANG Tao, LIU Yunjie, SUN Li
Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: To deal with the huge challenges in future mobile networks, the industry has proposed adaptive bit rate (ABR) technology and mobile edge computing (MEC), aiming to provide users with diverse services of high quality of experience, low latency and high bandwidth. Combining ABR and MEC to optimize the distribution of video content has been quite important for improving network performance and quality of experience. Especially, the joint optimization of network resources has arisen as an essential research topic. An overview of MEC was firstly given, and then the existing work in the industry on the joint optimization problem of MEC caching and transcoding oriented to adaptive streaming was analyzed and compared. Finally, the existing challenges in the future were summarized.

Key words: adaptive video streaming, MEC, caching, transcoding

1　引言

随着移动互联网技术的快速发展，移动网络数据流量呈现了爆发式的增长趋势，据 2017 年思科 VNI 报告，到 2021 年，全球移动数据流量将达到每月 49 EB（exabyte），其中移动视频流量将

收稿日期：2017−12−25；**修回日期**：2018−08−08
基金项目：中央高校基本科研业务费专项资金资助项目；国家自然科学基金资助项目（No.61501042）
Foundation Items：The Fundamental Research Funds for the Central Universities, The National Natural Science Foundation of China (No.61501042)

占全球移动数据流量的 78%[1]。爆发式增长的数据流量，尤其是视频流量给移动网络造成了巨大的挑战，要求移动网络具有提供更高数据传输速率和更低网络时延的能力，从而为用户提供更好的体验质量（quality of experience，QoE）。

为了应对上述网络挑战，一方面业界提出了在移动网络中采用自适应比特流（adaptive bit rate，ABR）传输技术进行视频分发[2]，即在移动网络中将视频文件编码为多种比特率版本，每个版本的视频文件都被切分成多个视频块（segment），并根据用户设备的能力、网络连接状况和特定的请求，为用户动态选择传送的视频块比特率版本，从而减少视频播放卡顿和重新缓冲率，提升用户的 QoE 体验[3]。另一方面，业界也提出了移动边缘计算（mobile edge computing，MEC），旨在移动网络边缘向内容提供商和应用开发者提供云计算能力和 IT 服务环境，从而为终端用户提供超低时延和高带宽的服务[4]。自 ABR 与 MEC 技术提出以后，通过利用 MEC 实现自适应视频流的缓存、转码与自适应分发受到了业界的广泛关注[5-6]，并表现出了显著的优势，主要体现在 MEC 在网络边缘提供了存储能力和计算能力，可实现视频内容的就近缓存，并根据动态变化的网络状况在网络边缘进行视频转码，以实现视频的自适应分发，从而降低视频内容的传输时延并节省网络带宽，同时提升用户的 QoE 体验。

目前已有大量工作将 MEC 与 ABR 技术联合考虑，并从不同的部署场景出发，针对视频流的缓存转码等问题进行了研究[7-18]，其中不仅包括单 MEC 服务器场景中的资源协同问题，还对分布式多 MEC 之间的资源协作问题进行了讨论，并针对不同的场景和需求，对缓存策略和转码策略进行了设计，从而优化包括系统成本、网络负载和视频容量在内的各项网络指标。另外，由于新型网络架构在增强视频业务质量方面展现的高效性能，部分研究工作还面向视频流分发业务提出了基于 SDN 的 MEC 架构，从而优化网络架构和视频分发机制。

尽管目前对 MEC 中缓存转码问题的研究已取得了一定的进展，但尚未有系统性的工作对基于 MEC 的自适应视频流分发研究的核心关键技术问题、业界已取得的进展以及未来仍面临的挑战等问题做全面的分析与总结。因此，本文将针对面向视频流的缓存转码联合优化问题，对目前已有工作进行总结和对比，并对目前仍需解决的问题进行归纳和展望。

2　自适应视频流场景下的 MEC 部署架构

欧洲电信标准化协会（European Telecommunications Standards Institute，ETSI）于 2014 年首次提出了移动边缘计算，并且给出了其“在移动网络边缘提供 IT 服务环境和云计算能力”的定义。MEC 将原本位于云数据中心的服务和功能下沉到网络边缘，提供超低时延、超高带宽的网络环境和实时网络分析能力。由于在回传网和核心网中发挥的低时延保证和容量增强的重要作用，MEC 技术受到了广泛关注，被公认为是 5G 的主要关键技术之一[19]。

ETSI MEC 基于虚拟化平台部署，目标是为第三方应用提供标准的体系架构和符合行业标准的 API，使得各种应用可以运行在网络边缘，同时基于 NFV（network function virtualization，网络功能虚拟化）的虚拟化基础设施有利于网络运营商的重复利用。MEC 服务器可以部署在网络边缘的不同位置，如 LTE 宏基站（eNode B）、3G 无线网络控制器（radio network controller，RNC）、汇聚点或核心网侧等[20]。

MEC 服务器通常具有较高的计算能力，适合于分析处理大量数据，可以为 AR 和 VR 等计算密集型业务提供高效的计算资源。另外，由于 MEC 在地理上十分接近用户或信息源，可以显著降低业务响应的时延，并减小回传网和核心网发

生网络拥塞的可能性。最后，位于接入网侧的MEC 能够实时感知网络数据，包括无线链路状况和用户行为信息及位置信息等，从而进行链路感知自适应，极大地改善用户的服务质量体验[21-22]。

与传统的网络架构相比，MEC 具有许多显著的优势，可以有效解决传统网络模式中高时延和低效率等问题。在面向视频流业务时，MEC 的存储计算资源以及网络感知能力可以有效地支持ABR 技术。一方面，MEC 的分布式边缘存储资源可以对视频内容进行缓存，并且对视频请求进行本地卸载，从而既可以缩短用户到视频内容的距离，降低传输时延，也可以减少视频内容的冗余传输，节省网络带宽并提高能量效率，缓解核心网络压力，另一方面，在 ABR 中，用户请求的视频比特率可基于网络条件、设备功能和用户偏好自适应调整[23]，因此往往需要对视频块进行多个比特率版本的缓存，而对一个视频块缓存多种比特率版本会造成较大的缓存成本，因此可以选择在 MEC 中缓存一部分较高比特率视频，对于缓存不命中请求，利用 MEC 的计算能力使用视频转码技术将缓存的视频版本转码为请求的比特率版本，视频转码可以提升缓存资源的利用率，在基于 ABR 的视频分发系统中具有重要的作用，此外，MEC 还可以利用其网络感知能力，对用户和网络信息进行分析，有效支持 ABR 技术。在自适应视频流业务中利用 MEC 缓存流行视频块，并在不同比特率版本之间进行转码，实现在网络边缘对用户请求进行响应，已经被认为是解决 ABR 中视频内容分发的一个重要趋势。

MEC 的缓存和转码能力可以有效支持 ABR 视频业务，从而缓解源服务器压力并提升视频观看质量，但在 MEC 资源容量有限的情况下，缓存转码资源的滥用会造成 MEC 负载过重，降低资源效率，影响用户体验，因此需要对不同视频业务的需求和特点进行分析，在各项资源之间进行权衡和分配，使得视频缓存转码的效率最优。除此之外，由于 MEC 分布式部署的特点，可以合理利用邻近 MEC 服务器的缓存转码资源，协作处理视频请求，优化整个网络的视频分发质量和资源效率。因此，研究 MEC 中面向视频流的计算、缓存与网络资源的联合优化问题，均衡地对缓存、计算和带宽资源进行合理编排，对优化系统整体性能具有重要意义。

图 1 为一个自适应视频流场景下的 MEC 部署架构，MEC 服务器的位置部署在基站之后，其中部署了缓存和计算资源，可以对视频文件进行缓存和转码，当用户的请求到达 MEC 时，若缓存命中或转码命中，就可以在本地对请求进行响应，如图 1 中的(a)路线和(b)路线，而不需要经过图 1 中(c)路线所示的回传网和核心网的传输，从源服务器获取视频文件。这样一来，既显著降低了用户请求响应的时延，保证用户体验，同时还可以避免网络拥塞，节省回传网和核心网的资源。

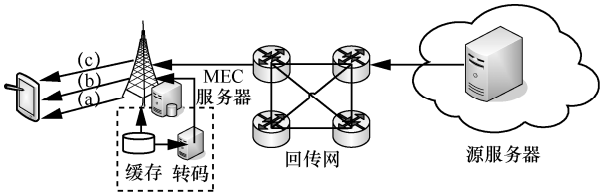


图 1　自适应视频流场景下的 MEC 部署架构

3　面向视频流的缓存转码资源联合优化

利用 MEC 实现 ABR 技术，可以有效解决视频内容的缓存转码和分发问题，其中缓存转码资源的联合优化已经被认为是一个重要的研究课题。在传统的视频流分发系统中，缓存和转码都在云端完成，由于传统方案中较高的传输时延和回传压力，MEC 逐渐得到了广泛的应用，研究结果表明部署 MEC 可以有效提升系统性能和用户体验。

3.1　在云端的缓存转码联合优化

媒体云作为当前一种有效实现自适应流分发的框架，在基于云的体系架构中，动态按需编排

虚拟化存储和计算资源，为显著降低系统成本提供了可能性，图 2 为在媒体云中应用自适应视频流业务的缓存转码示意图，媒体云的流媒体引擎从媒体库中获取视频内容，并进行视频内容的缓存和转码，进而实现自适应视频流向用户的分发。Jin 等人[7]研究了媒体云中缓存和转码的最优化策略问题，旨在动态调度各种资源，从而最小化某个节点响应单个请求的总运营成本。具体来说，在缓存策略方面，该方案考虑了视频的流行度，即对流行度最高的若干视频内容缓存全部比特率版本，对流行度次高的一部分视频内容缓存最高比特率版本。当在节点缓存命中时，直接利用缓存响应请求；当转码命中时，比较转码成本和带宽成本，选择成本较小的方式提供服务；当本地缓存不可用时，则从源服务器获取内容。参考文献[7]将缓存、转码和带宽成本之间的权衡问题建模成一个凸优化问题，然后采用两步分析法，分别对缓存资源分配和转码配置策略进行优化，最后得到最优的节点缓存空间大小和所有版本都缓存的视频数量。

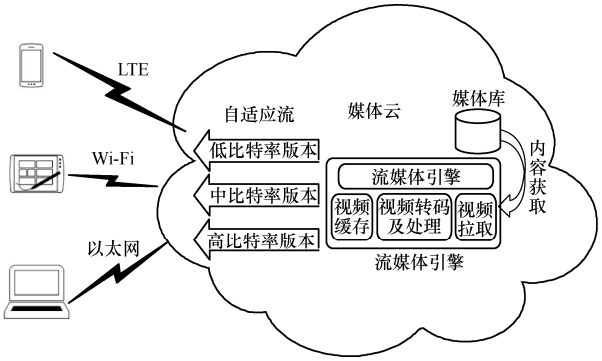


图 2　媒体云中面向自适应视频流的缓存与转码

面向视频流的资源联合优化，不仅可以从网络资源配置的角度出发，还可以从用户请求内容的角度进行视频块的调度决策。Gao 等人[8]根据对用户视频行为的分析，研究了媒体云中视频内容的管理问题，提出了一种成本高效的部分转码方案。具体来说，每个视频被分为一组片段，其中一部分片段被预处理并存储在缓存中，而其他片段则在播放过程中被快速转码。缓存策略方面，服务器中存储着源视频文件和一部分转码版本，源文件可以转码为任何比特率的版本，另外基于用户实时变化的比特率需求和部分转码的方案，采取动态缓存策略，从而在每个时隙根据用户请求信息来决定是否缓存。当用户请求在节点缓存命中时，用户可以直接获取缓存中的视频内容，进而产生存储成本，缓存不命中时，则进行在线转码，进而产生计算成本。该方案考虑了与存储和计算相关的总成本，建立了一个约束随机优化问题，优化目标为最小化某个服务器的长期总成本，决定某个视频段应缓存还是在线转码，然后利用李雅普诺夫优化框架和拉格朗日松弛法，设计了在线算法进行求解。

通过对用户视频行为的分析可以看出，超过 60%的视频只有前 20%的内容被用户观看，视频内部存在着不同的流行度，从这一角度出发可以进一步优化网络资源调度。Zhao 等人[9]研究了云网络中缓存和转码的均衡问题，基于对视频的分段，考虑了视频段之间的不同流行度以及云中存储和计算价格的大小，对系统中所有视频的放置问题进行了研究，从而决定对每个视频段的哪些版本进行缓存或者转码，将总的系统运营成本最小化。具体来说，首先根据视频内部流行度将视频文件分为多个视频段，通过一个转码权重图描述不同比特率版本之间的转码关系，从而计算出各版本之间的转码成本。基于以上的考虑，提出了一个存储和转码的权衡策略，对流行视频段存储多个或者所有版本，对不流行的视频段存储最高比特率版本，并针对用户请求进行转码。为了避免转码启动时延对视频播放效果的影响，可以在播放视频段时，提前对下一个视频段进行转码。该方案将该权衡问题描述为一个优化问题，优化目标为最小化与存储和转码相关的总成本，并利用启发式分治算法求解，进行视频段某个版本的

缓存和转码决策。

3.2 在 MEC 中的缓存转码联合优化

在云端进行缓存和转码，会带来较高的传输时延，增加回传网和核心网的带宽压力，而 MEC 逐步展现出应对视频业务挑战的优势，利用 MEC 的存储和计算能力，可以有效提高网络的整体性能，并显著改善用户体验。其中，MEC 中缓存和转码的联合优化问题已成为一个重要的研究课题。

图 3 所示为在分布式 MEC 架构中实现自适应视频流业务的示意，其中，每个 MEC 服务器都可以实现视频的缓存和转码，分布式部署的 MEC 服务器之间可以实现协作式的缓存和转码，从而实现资源的高效利用，进一步提升视频分发效率和用户体验。

为解决无线网络中自适应比特流的缓存挑战，Pedersen 等人在参考文献[10-11]中研究了自适应比特流场景中无线缓存和处理的联合优化问题。该方案首先将视频文件分为多个视频块，每个视频块可以按不同的比特率请求。针对 RAN 的缓存挑战，提出在 RAN 部署有限的计算资源，从而可以进行视频块之间的转码，缓解存储压力。基于以上考虑，提出了基于 ABR 感知的主动/被动的联合转码和缓存资源的策略，具体来说，对于视频请求有 3 种内容获取方式，从缓存处直接获取对应版本，对缓存的高比特率版本进行转码，或者通过回传网络从 CDN 处获取。当用户请求在节点缓存命中时，直接由缓存进行响应，缓存不命中时，可以根据给定可用的缓存容量、处理能力和回传带宽，通过转码资源和回传资源分配算法进行转码和回传决策，当采用回传方式时，采用缓存策略对获取的内容进行缓存。参考文献[10-11]中分别采用了两种缓存方式，分别是 LRU（least recently used）缓存策略和 P-UPP（proactive user preference profile）缓存策略。该方案制定了一个优化问题，优化目标为最大化无线网络的视频容量，即服务的并发视频请求数量，并采用启发式算法进行了求解，从而对某个视频块的获取方式进行调度决策。

Wang 等人[12-13]提出了一个在线转码和地域分布式交付的联合策略，系统架构中包括多个 CDN 区域，每个区域中包含后端服务器和对等服务器，转码任务在后端服务器中完成。该方案考虑了用户的 CDN 区域偏好、区域的转码版本偏好以及视频请求的用户偏好。首先根据用户的 CDN 区域偏好，即考虑服务器到用户的带宽大小对用户进行重定向，选择提供服务的 CDN 区域，该区域中的对等服务器以循环方式提供服务。另外，根据内容的用户偏好和区域的转码版本偏好来安排转码任务，并选择空闲的 CDN 计算资源进行转

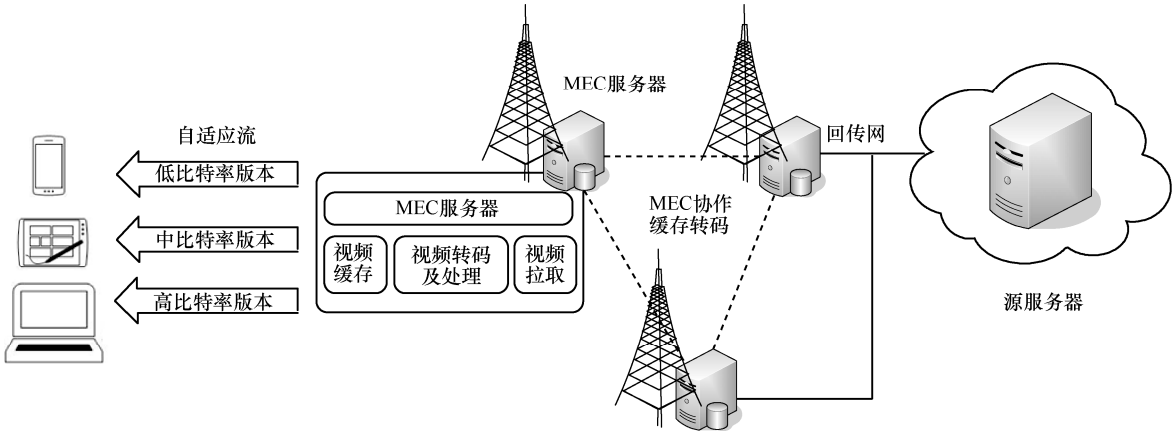


图 3 分布式 MEC 架构中面向自适应视频流的协作缓存与转码

码和交付，减少跨区域的复制成本，特别地，根据按需设计的策略，视频段被转码为一组预定义的版本，如果转码不及时，可以选择最接近的比特率版本进行转发。最后，该方案对优化问题进行了建模，优化目标为最小化计算成本和复制成本，并采用启发式和分布式算法进行了求解。

在视频业务中，当视频比特率和传输条件不匹配时，会引发网络的拥塞和较高的时延，造成视频播放的卡顿，严重影响用户的观看体验，因此，在资源联合优化的过程中，不仅要从运营商的角度出发考虑成本代价，还要从用户的角度出发保障 QoE。众包直播游戏视频流（crowdsourced live game video streaming，CLGVS）是一种新兴的互联网业务，可以使众多异构终端随时随地观看游戏玩家播放的视频，Zheng 等人[14]研究了 CLGVS 业务中的在线转码和交付问题，从 CLGVS 服务提供商的角度，通过联合优化动态转码决策、比特率配置和数据中心选择，减少运营成本并保证用户的服务质量。该方案考虑了两种转码策略，分别为门限转码策略和全部转码策略。该问题被建模为一个约束随机优化问题，优化目标包括两部分，即最小化与计算和带宽相关的总运营成本，最大化与时延和比特率相关的用户 QoE。然后利用李雅普诺夫优化框架，设计了一个在线算法 OCTAD（online cloud transcoding and distribution）进行求解，算法包括 3 个主要部分：动态直播转码决策、自适应比特流配置和智能数据中心选择，从而动态地为每个游戏玩家执行比特率配置，为每个观看者进行转码决策和数据中心选择。同时，为了扩展算法的适用范围，还在设计在线算法时考虑了游戏的类型。

现有的缓存转码优化方案主要基于单服务器独立进行缓存和转码任务决策的场景，没有考虑服务器之间的协作，为了研究在多 MEC 服务器场景下联合优化整体运营成本的问题，Tran 等人[15-16]提出了一种移动边缘计算网络中多比特率视频流的协作缓存和处理策略，称为 CoPro-CoCache。在该方案中，获取视频内容的方法可能有：从本地服务器的缓存中获取，在本地服务器中转码获得，从协作服务器的缓存中获取，在协作服务器中进行转码并传回，从协作服务器传回后在本地进行转码。具体来说，缓存策略方面，本方案不需要内容流行度的先验信息，采用 LRU 缓存策略，将每个小区最流行的视频缓存在对应的基站缓存服务器上，直到缓存存储空间已满，当用户的视频请求需要对缓存中的比特率版本进行转码时，将转码任务分配给负载最小的 MEC 服务器，从而均衡网络负载，这个服务器可以是存储原始版本的 MEC 服务器，即数据提供节点，也可以是交付节点。参考文献[16]将该协作缓存和处理问题建模为一个整数线性规划问题，该问题受存储空间和处理能力的约束，在给定可用资源后，优化目标为对单个视频请求协作制定缓存放置策略和视频调度策略，从而最小化回传网络成本。最后，针对该 NP 问题提出了一个新型的在线算法 JCCP（joint collaborative caching and processing）来进行求解。

Xu 等人[17]提出了 MEC 增强的自适应比特率（MEC-ABR）视频传输方案，联合进行缓存和无线资源的分配。在该方案中，MEC 服务器作为控制组件来执行缓存策略并灵活调整视频版本。具体来说，该方案首先考虑了 BS 的流量负载，从而进行 MEC 服务器的存储资源分配，以缓存各 BS 服务范围内的流行视频，并将该存储资源分配问题模拟为 Stackelberg 博弈进行了求解。缓存策略方面，不仅考虑了视频的流行度，还考虑了 RAN 侧的无线信道质量，缓存策略和视频交付可以被灵活地调整，以匹配不同的无线信道。另一方面，该方案将联合缓存和无线资源的分配问题建模为匹配问题，BS 和用户分别根据视频的流行度和无线信道条件维护偏好列表，利用 MEC 的存储和计算能力进行优化，提出了 JCRA（joint cache and

radio resource allocation）算法来解决这个问题，并考虑了视频比特率版本的动态调整。

软件定义移动网络（software-defined mobile network，SDMN）、网内缓存和 MEC 作为下一代移动网络的重要技术，对增强视频业务质量具有重要意义，Liang 等人[18]研究了一个 MEC-SDMN 中的视频速率适应问题，联合考虑视频速率自适应、带宽配置和 MEC 中的计算资源调度，设计了一个高效机制。研究目的是在考虑网络资源和视频缓存分布的情况下，为每个用户找到最佳的视频质量水平。在该方案中，SDN 控制器执行流量管理，通过最大限度地增大视频的整体平均增益，提高整个网络的效用，以帮助用户自适应地选择最佳的视频质量水平。为了最大化 HetNet 的平均视频质量，该方案制定了一个优化问题，并采用双分解方法，将视频数据速率、计算资源和流量管理（带宽配置和路径选择）3 部分问题解耦，独立求解各个变量。

3.3 方案对比

本章前两部分主要从云端和 MEC 两个场景出发，介绍了目前已有工作中面向视频流的缓存、计算和带宽资源的联合优化方案。针对以上方案策略和优化方法等的不同，下面从缓存策略、转码策略、建模方法和算法等具体方面对已有方案进行了对比，见表 1。

从表 1 可以看出，已有工作主要面向的是单服务器架构，对协作缓存转码问题考虑得较少。缓存方面主要采用一部分全部缓存和一部分缓存最高版本的策略，在考虑视频流行度的情况下，可以缓存流行视频的全部版本，而对不流行视频缓存最高比特率版本。转码策略包括以下几种策略：缓存不命中直接转码，与带宽资源做均衡进行转码决策，协作转码中考虑负载和成本进行转码决策等。另外，建立优化问题时，优化目标主要为最小化系统成本，除此之外，还包括对视频容量和 QoE 等方面的优化。解决问题的算法主要包括李雅普诺夫优化理论、分析法、博弈论、启发式算法和在线算法等。

4 面向视频流的 MEC 资源优化问题与挑战

在网络边缘部署 MEC 来应对视频流业务，可以有效降低传输时延并节省网络资源。虽然目前已有大量工作对 MEC 面向 ABR 的缓存、计算和带宽资源的分配和调度问题进行了研究，但如何综合考虑网络各方面因素，均衡各项资源，从而使系统整体性能最优，有多方面的挑战和研究难点，本文总结了以下 5 个方面。

4.1 缓存转码带宽资源优化与 QoE 优化的均衡问题

QoE 是一种以用户认可程度为标准的服务评价方法，直接反映了用户在一定客观环境中对适用的服务或业务的整体认可程度[24]。自适应流媒体的发展是推动探索增强 QoE 的有效方法的关键驱动力，从而通过对用户提供差异化服务来保障用户体验[25]。MEC 中缓存转码带宽资源的优化与 QoE 优化的均衡问题是一个非常有意义的研究方向，从视频内容提供商的角度出发，对于系统的优化一般要考虑两个维度：一方面要降低缓存、计算和网络的运营成本，另一方面又要保证终端用户的 QoE。因此，如何权衡 MEC 缓存转码带宽资源的租赁成本与终端用户的 QoE 保证，是今后研究的一个重要方向。

4.2 缓存转码带宽资源的能量效率优化问题

MEC 的部署将原本位于云端的存储和计算资源下沉到网络边缘，一方面使得网络边缘可以对用户请求进行响应，一方面可以减少回传资源的浪费。在 MEC 的网络优化方面，能量效率问题是关注的重点问题之一。在 MEC 的部署场景中，内容的缓存、MEC 的计算以及 MEC 之间、MEC 与用户之间的通信都会产生大量的能耗，从而带来极大的能耗成本。因此，建立能量高效的资源优化机制，对缓存、计算和通信资源进行有效的调度，对于减少系统能耗、提高系统性能有着重

表 1　缓存转码联合优化方案对比

	参考文献	架构	缓存策略	转码策略	建模	解决问题	算法	偏好考虑
云端	[7]	单服务器	缓存最流行的视频的所有版本，缓存不流行视频的最高版本	对缓存不命中，有最高版本缓存，且转码成本低于带宽成本的请求进行转码	凸优化问题 优化目标：最小化与缓存、计算和带宽相关的总运营成本	资源配置：服务器缓存资源分配和转码策略	两步分析法	视频流行度
	[8]	单服务器	缓存最高版本的源视频文件以及一部分转码版本	部分转码	约束随机优化问题 优化目标：最小化与存储和计算相关的长期总成本	某视频段的缓存和转码策略	利用李雅普诺夫优化框架和拉格朗日松弛法设计的在线算法	无
	[9]	单服务器	流行视频块缓存多个版本，不流行视频块缓存最高比特率版本	缓存不命中时进行转码，还可以提前对下一个视频段进行转码	最小化存储和转码相关的总成本	所有视频段的缓存和转码策略	启发式分治算法	视频段流行度
MEC	[10] [11]	单服务器	LRU 缓存策略 P-UPP 缓存策略	通过转码资源和回传资源分配算法进行转码决策	多背包问题 优化目标：最大化网络视频容量和 QoE	某视频块的调度策略	启发式算法	无
	[12] [13]	多区域多服务器协作	缓存视频块的最高比特率版本	协作转码	优化目标：减少计算资源消耗，最小化复制成本	转码资源的调度	启发式和分布式算法	请求用户偏好、区域用户偏好和转码版本区域偏好
	[14]	多服务器无协作	缓存最高比特率版本的视频	门限转码策略和全部转码策略	约束随机优化问题 优化目标：最小化与计算和带宽相关的总运营成本，最大化与时延等相关的用户 QoE	为每个游戏玩家执行比特率配置，为每个观看者进行转码决策和数据中心选择	利用李雅普诺夫优化框架设计的在线算法 OCTAD	无
	[15] [16]	多服务器协作	缓存最流行的视频块，采用 LRU 缓存策略，协作缓存	协作转码，选择负载最小的服务器进行转码	整数线性规划问题 优化目标：最小化回传网络成本，受缓存和处理能力约束	某视频块的缓存放置策略和调度策略	JCCP 在线算法	无
	[17]	单服务器	缓存各小区流行视频	按需转码	MEC 缓存资源分配：Stackelberg 博弈 联合缓存和无线资源分配：匹配问题	资源分配问题：缓存资源和无线资源	Stackelberg 博弈论；JCRA 在线算法	视频流行度
	[18]	多服务器	缓存一部分视频的最高版本	根据计算容量进行转码决策	优化目标：最大化平均视频质量水平	为每个用户寻找最佳视频质量	双分解方法	无

要意义。在面向视频流的 MEC 缓存转码带宽资源的联合优化中，主要关注缓存能耗、转码能耗和传输能耗，如何联合考虑 MEC 的计算、转码和传输，优化提供视频流业务的能量效率是今后研究的一个重点。

4.3　MEC 中基于深度增强学习的缓存和转码

深度增强学习是将深度学习和增强学习结合起来，从而实现端对端学习的一种全新的算法，是通用的人工智能框架，目前已经成为网络优化的重要方法和工具[26]。结合深度增强学习的方法，对自适应视频流内容进行缓存是一个重要研究方向。在基于 ABR 视频流缓存系统中，每个视频块都有多个比特率版本，考虑到边缘网络缓存系统的容量限制，缓存所有比特率的视频块会造成缓

存资源利用率的降低和网络成本的增加。通过部署在网络无线接入侧的 MEC，可以实时对网络信息进行感知，包括网络链路状况和用户行为等数据[22]，利用深度增强学习的方法对这些信息进行分析和学习，可以预测视频内容的流行度以及用户对响应视频块比特率版本的请求状况，提前进行资源分配和调度，对相应视频内容和比特率版本进行缓存，从而提高缓存命中率和缓存资源的利用率。

4.4 分布式的多 MEC 协作问题

MEC 在边缘网络中的部署，通常采用分布式的方式，因此 MEC 带来的缓存和计算资源也分布式的位于网络的不同位置。单个 MEC 的存储空间和计算能力都是有限的，过多的缓存和计算任务会给 MEC 服务器造成过载，而回传到云数据中心又会产生较高的回传成本，因此基于 MEC 分布式的部署方式，相邻的 MEC 服务器之间可以协作进行缓存和计算，当前 MEC 服务器没有相应缓存内容或计算资源紧张时，可以调用其他空闲 MEC 服务器，此类分布式协作的方式可以有效减少网络运营成本，提高网络性能[27]。因此，不同 MEC 节点之间如何协作共享资源（主要包括计算和缓存资源）成为一个重要的研究问题。例如，当用户请求的目标视频内容在本地 MEC 服务器没有缓存时，如何在其他缓存有相应内容的 MEC 节点中选择一个最优的节点；当本地 MEC 服务器的计算负荷过载时，如何将本地的计算任务卸载至其他的 MEC 节点，这都需要 MEC 节点之间的协作。因此研究基于分布式的多 MEC 协作的资源共享机制，以提高资源的利用率和用户的体验也是今后进行资源联合优化的一个重要方向。

4.5 基于 NFV、SDN 和网络切片等新型技术的资源分配问题

在自适应视频流场景中，不同的网络环境和用户能力可以动态适配视频流的比特率版本，而不同类型的视频业务和不同等级的用户对 QoE 保障有着差异化的需求，另一方面，由于单 MEC 服务器的计算和存储资源有限，分布式多 MEC 场景下的资源协同也面临着很大的需求和挑战[28]。如何针对不同业务场景，利用新型网络技术实现资源的高效管理和分配是 MEC 中面向视频流业务的重要研究课题。网络切片技术可以针对不同应用场景，将物理网络切割成多个虚拟网络，从而应对不同场景中对传输时延、移动性、可靠性、安全性以及计费方式的差异性，利用边缘计算的计算、存储和通信能力，构建业务所在无线接入网络内的接入网切片，可以实现业务的本地处理，缓解核心网压力，减少传输时延，改善业务性能[29]。除此之外，未来的 5G 网络还提出了如下演进目标：基于 SDN/NFV 进行虚拟化，进行扁平化扩展与增强，其中 NFV 和 SDN 是实现网络切片的基础，NFV 提供了按需分配的可配置资源共享池，可以极大地方便资源的统一管理，同时 SDN 实现了集中式的控制平面，并通过为用户提供的编程接口，使用户可以根据上层业务和应用个性化地定制网络资源来满足其特有的需求[30]。针对不同的业务场景，进行有效的网络业务切片和划分，并通过 SDN 的全局管控和对 NFV 虚拟资源的合理调配，对优化整体资源效率和网络性能具有重要研究意义。

5 结束语

本文从 MEC 和 ABR 的背景和概述出发，对目前面向视频流的缓存转码资源联合优化方案进行了介绍和分析，并主要从缓存策略、转码策略和优化方式等方面对已有方案进行了对比。在对以上方案分析对比的基础上，研究了面向视频流的 MEC 资源优化问题目前面临的挑战和研究难点，如与时延优化的均衡问题、能量优化问题和用户行为分析以及分布式 MEC 协作的问题等，在网络整体优化方面具有重要意义。

参考文献：

[1] Cisco Mobile VNI. Cisco visual networking index: global mobile data traffic forecast update, 2016–2021 white paper[R]. 2017.

[2] STOCKHAMMER T. Dynamic adaptive streaming over HTTP: standards and design principles[C]//The Second Annual ACM Conference on Multimedia Systems, Feb 23-25, 2011, San Jose, CA, USA. New York: ACM Press, 2011: 133-144.

[3] YIN X Q, JINDAL A, SEKAR V, et al. A control-theoretic approach for dynamic adaptive video streaming over HTTP[C]//The 2015 ACM Conference on Special Interest Group on Data Communication Pages, August 17-21, 2015, London, UK. New York: ACM Press, 2015: 325-338.

[4] ETSI. Mobile edge computing—a key technology towards 5G[R]. 2015.

[5] LI Y, FRANGOUDIS P A, HADJADJ-AOUL Y, et al. A mobile edge computing-based architecture for improved adaptive HTTP video delivery[C]//2016 IEEE Conference on Standards for Communications and Networking (CSCN), Oct 29-31, 2016, Paris, France. Piscataway: IEEE Press, 2016: 1-6.

[6] WANG C C, LIN Z N, YANG S R, et al. Mobile edge computing-enabled channel-aware video streaming for 4G LTE[C]//Wireless Communications and Mobile Computing Conference, June 26-30, 2017, Valencia, Spain. Piscataway: IEEE Press, 2017: 564-569.

[7] JIN Y, WEN Y, WESTPHAL C. Optimal transcoding and caching for adaptive streaming in media cloud: an analytical approach[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(12): 1914-1925.

[8] GAO G, ZHANG W, WEN Y, et al. Towards cost-efficient video transcoding in media cloud: insights learned from user viewing patterns[J]. IEEE Transactions on Multimedia, 2015, 17(8): 1286-1296.

[9] ZHAO H, ZHENG Q, ZHANG W, et al. A segment-based storage and transcoding trade-off strategy for multi-version VoD systems in the cloud[J]. IEEE Transactions on Multimedia, 2017, 19(1): 149-159.

[10] AHLEHAGH H, DEY S. Adaptive bit rate capable video caching and scheduling[C]//IEEE WCNC'13, April 7-10, 2013, Shanghai, China. Piscataway: IEEE Press, 2013: 1357-1362.

[11] PEDERSEN H A, DEY S. Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing[J]. IEEE/ACM Transactions on Networking, 2016, 24(2): 996-1010.

[12] WANG Z, SUN L, WU C, et al. Joint online transcoding and geo-distributed delivery for dynamic adaptive streaming[C]//IEEE INFOCOM'14, April 29-May 1, 2014, Toronto, Canada. Piscataway: IEEE Press, 2014: 91-99.

[13] WANG Z, SUN L, WU C, et al. A joint online transcoding and delivery approach for dynamic adaptive streaming[J]. IEEE Transactions on Multimedia, 2015, 17(6): 867-879.

[14] ZHENG Y, WU D, KE Y, et al. Online cloud transcoding and distribution for crowdsourced live game video streaming[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(8): 1777-1789.

[15] TRAN T X, HAJISAMI A, PANDEY P, et al. Collaborative mobile edge computing in 5G networks: new paradigms, scenarios, and challenges[J]. IEEE Communications Magazine, 2017, 55(4): 54-61.

[16] TRAN T X, PANDEY P, HAJISAMI A, et al. Collaborative multi-bitrate video caching and processing in mobile-edge computing networks[C]//WONS'13, March 18-20, 2017, Banff, Canada. [S.l.:s.n.], 2017: 165-172.

[17] XU X, LIU J, TAO X. Mobile edge computing enhanced adaptive bitrate video delivery with joint cache and radio resource allocation[J]. IEEE Access, 2017(5): 16406-16415.

[18] LIANG C, HU S. Dynamic video streaming in caching-enabled wireless mobile networks[J]. arXiv: 1706.09536, 2017.

[19] 王胡成, 徐晖, 程志密,等. 5G 网络技术研究现状和发展趋势[J]. 电信科学, 2015, 31(9): 149-155.
WANG H C, XU H, CHENG Z M, et al. Current research and development trend of 5G network technologies [J]. Telecommunications Science, 2015, 31(9): 149-155.

[20] 李子姝, 谢人超, 孙礼, 等. 移动边缘计算综述[J]. 电信科学, 2018, 34(1): 87-101.
LI Z S, XIE R C, SUN L, et al. A survey of mobile edge computing[J]. Telecommunications Science, 2018, 34(1): 87-101.

[21] TALEB T, SAMDANIS K, MADA B, et al. On multi-access edge computing: a survey of the emerging 5G network edge cloud architecture and orchestration[J]. IEEE Communication Surveys & Tutorials, 2017, 19(3): 1657-1681.

[22] WANG S, ZHANG X, ZHANG Y, et al. A survey on mobile edge networks: convergence of computing, caching and communications[J]. IEEE Access, 2017, 5(3): 6757-6779.

[23] TIMMERER C, GRIWODZ C. Dynamic adaptive streaming over HTTP: from content creation to consumption[C]//Proc. of ACM MM'12, Oct 29-Nov 2, 2012, Nara, Japan. New York: ACM Press, 2012: 1533-1534.

[24] 赵希鹏，张欣，杨大成，等. 基于 QoE 的无线网络资源调度优化研究[J]. 移动通信, 2014(22): 8-13.
ZHAO X P, ZHANG X, YANG D C, et al. Research on optimization of wireless network resource scheduling base on QoE[J]. Mobile Communications, 2014(22): 8-13.

[25] LI C, TONI L, ZOU J, XIONG H, et al. QoE-driven mobile edge caching placement for adaptive video streaming[J]. IEEE Transactions on Multimedia, 2017(9): 1.

[26] HE T Y, ZHAO N, YIN H. Integrated networking, caching and computing for connected vehicles: a deep reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2017, 99(10): 1.

[27] GHARAIBEH A, KHREISHAH A, JI B, et al. A provably efficient online collaborative caching algorithm for multicell-coordinated systems[J]. IEEE Transactions on Mobile Computing, 2016, 15(8): 1863-1876.

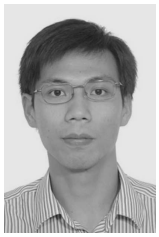
[28] WANG W, CAO J, ZHANG W. Edge computing: vision and challenges[J]. IEEE Internet of Things Journal, 2016, 3(5): 637-646.

[29] 项弘禹, 肖扬文, 张贤, 等. 5G 边缘计算和网络切片技术[J].

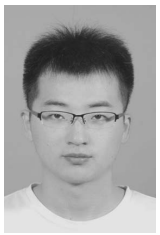
电信科学, 2017, 33(6): 54-63.
XIANG H Y, XIAO Y W, ZHANG X, et al. Edge computing and network slicing technology in 5G[J]. Telecommunications Science, 2017, 33(6): 54-63.
[30] GPPP E B. QoE-oriented mobile edge service management leveraging SDN and NFV[J]. Mobile Information Systems, 2017(1).

[作者简介]

李佳（1994−），女，北京邮电大学未来网络理论与应用实验室硕士生，主要研究方向为 5G 网络、移动边缘计算等。



谢人超（1984−），男，北京邮电大学未来网络理论与应用实验室副教授、硕士生导师，主要研究方向为信息中心网络、移动网络内容分发技术和移动边缘计算等。



贾庆民（1990−），男，北京邮电大学未来网络理论与应用实验室博士生，主要研究方向为新型网络体系架构、内容分发和移动边缘计算等。

黄韬（1990−），男，北京邮电大学未来网络理论与应用实验室教授、博士生导师，主要研究方向为新型网络体系架构、内容分发网络软件定义网络等。

刘韵洁（1943−），男，中国工程院院士，北京邮电大学教授、博士生导师，主要研究方向为未来网络体系架构。

孙礼（1959−），男，北京邮电大学未来网络理论与应用实验室副教授、硕士生导师，主要研究方向为宽带通信网络、无线接入技术、通信网络交换技术等。

研究与开发

工业物联网无线信道与噪声特性

张克[1]，刘留[1,2]，袁泽[1]，张琨[1]，张建华[2]，刘志军[3]
（1. 北京交通大学电子信息工程学院，北京 100044;
2. 北京邮电大学泛网无线通信教育部重点实验室，北京 100876;
3. 北京航天测控技术有限公司，北京 100041）

摘　要：随着“中国制造 2025”“智能制造”“互联网+”等一系列国家战略规划的提出和实施，国内工业物联网技术将迎来迅猛的发展。然而，工厂恶劣环境下的信道和噪声特性给工业物联网无线通信带来了极大的挑战。基于此，对工业物联网应用的无线通信技术进行介绍和对比，总结了工业物联网环境下的信道和噪声特点，对工业物联网无线通信研究中的关键内容——信道和噪声特性分析进行了回顾。

关键词：工业物联网；无线通信技术；信道特性；噪声特性
中图分类号：TN929.5
文献标识码：A
doi：10.11959/j.issn.1000−0801.2018217

Wireless channel and noise characteristics in industrial internet of things

ZHANG Ke[1], LIU Liu[1,2], YUAN Ze[1], ZHANG Kun[1], ZHANG Jianhua[2], LIU Zhijun[3]
1. Institute of Broadband Wireless Mobile Communications, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
2. Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
3. Beijing Aerospace Measurement and Control Technology Co., Ltd., Beijing 100041, China

Abstract: With the proposal and implementation of “Made in China 2025” “Intelligent Manufacturing” “Internet plus” and a series of national strategic planning, the industrial internet of things (IIoT) has developed rapidly in China. However, the channel and noise characteristics of the industrial poor environment have brought great challenges to the wireless communication of the industrial internet of things. The application of different wireless communication technologies in IIoT were introduced and compared and the features of wireless channel and noise in the environment of IIoT were summarized. As the key content of wireless communication research, channel and noise characteristics were analyzed overall.

Key words: industrial internet of things, wireless communication technology, channel characteristics, noise characteristics

收稿日期：2018−01−20；修回日期：2018−05−20
基金项目：北京邮电大学泛网无线通信教育部重点实验室资助项目（No.KFKT-2018105）；北京市科技新星计划基金资助项目（No.Z161100004916068）；国家自然科学基金面上基金资助项目（No.61471027）
Foundation Items: Key Laboratory of Universal Wireless Communications, Ministry of Education (No.KFKT-2018105), Beijing New-star Plan of Science and Technology (No.Z161100004916068), The National Natural Science Foundation of China (No.61471027)

1 引言

随着工业化与信息化的深度融合，企业内部互联互通的需求渐增，通过接入网络进而达到提高产品质量和运营效率的需求更为强烈，IIoT（industrial internet of things，工业物联网）应运而生。由中国电子技术标准化研究院编写的《工业物联网白皮书（2017）》指出：“工业物联网是通过工业资源的网络互联、数据互通和系统互操作，实现制造原料的灵活配置、制造过程的按需执行、制造工艺的合理优化和制造环境的快速适应，达到资源的高效利用，从而构建服务驱动型的新工业体系”[1]。目前，提高生产效率、实现节能减排是我国制造业面临的主要战略任务，伴随着工业物联网的发展，智能制造将贯穿于企业生产经营的各个环节，为我国制造业的发展带来深刻的变革。2017 年 1 月由工业和信息化部发布的《物联网的十三五规划（2016—2020 年）》提出，建设制造强国、网络强国，推进供给侧结构性改革，以 CPS（cyber-physical system，信息物理系统）为代表的物联网智能信息技术将在制造业智能化、网络化、服务化等转型升级方面发挥重要作用[2]。由此可见，制造业已经成为工业物联网的重要应用领域。

无线通信技术是工业物联网发展的重要基础，工业物联网的实施一般包括 4 个阶段：一是利用智能感知技术随时随地采集工业数据，二是通过通信网络将采集的数据传递出去，三是利用云计算、大数据等技术对这些数据进行深度挖掘和利用，四是基于信息管理、智能终端和平台集成等技术，实现传统工业的智能化改造[1]。正是通信技术的发展保证了第二阶段的顺利进行，通信网络连接现场设备、控制器、人机界面、监控系统以及企业管理系统，是工业物联网生产系统中的信息传输通道，是生产系统稳定安全运行的重要基础。而在工业通信网络中，无线通信是其中重要的组成部分之一，相较于有线通信网络，无线通信网络构建成本低，减少了大量电缆安装、维护所需的费用和时间，避免了振动、高温等恶劣环境对电缆的损坏。工业物联网中的无线通信技术主要可以分为两类：一类是 ZigBee、Wi-Fi、Z-wave、蓝牙（bluetooth）等短距离通信技术；另一类是 LPWAN（low-power wide-area network，低功耗广域网络），比较常见的如 NB-IoT（narrow band internet of things，基于蜂窝网络的窄带物联网）、eMTC（增强机器类通信）、LoRa（基于扩频技术的超远距离无线传输方案）。业界对于 5G 提出了需求各异的应用场景，目前 5G 系统已经包括工业环境的通信概念，涵盖广泛的应用，包括机器类型通信、移动网络物理系统和智能工厂，这些将会使工业自动化界获益匪浅[3]。同时，mMTC（massive machine type communication，大规模机器类通信）是 ITU-R（ITU-Radio Communications Sector，国际电信联盟无线电通信组标准化组织）确定的 5G 三大主要应用场景之一，海量的无线物联网的研发和应用必将有效地支撑工业物联网无线技术的发展[4]。

从信息论的角度看，决定通信频谱效率有两个因素：一是传输的信道特征，二是传输链路的信噪比。从信道特征来看，工业场景中的传播环境与传统无线通信的传播环境存在较大的差别，传统工业有煤炭厂、钢铁厂、机械厂、木材厂、服装厂等，不同的产业制造和生产不同的产品，这些材料吸收和反射能力有很大差异，其信道传播特征（包括多径分量、路径损耗等）均表现不同；此外，工业场景中存在的机床、机械臂等金属障碍物会对电波传输损耗造成影响；金属设备在电波传播中会形成较强的镜面反射和散射，从而产生更多强度较大的多径分量；工业自动化中的机械臂转动、机器人运输移动等运动因素会让无线信道同时具有时变特性。从传输链路信噪比来看，在常规无线通信中信噪比的定量使用中，

通常使用加性高斯白噪声，即噪声的功率谱是一个常数。工厂在工作时，由于设备温度的升高、机械震动、火花放电等物理现象会辐射出大量的电磁噪声，这时会出现突发的脉冲噪声，这些噪声可能在功率谱形状、生命周期等方面和传统的加性高斯白噪声有着较大的不同；同时，工业物联网技术需要大量应用低功耗无线传感设备，工厂中的电磁噪声会对低功耗的无线传感系统产生巨大的影响。

无线通信想要真正在工业物联网发挥作用，需要对工业环境下无线信道和噪声特性开展新的研究分析，因为无线通信系统的传输速率和质量最终都要受到无线信道和噪声特性的制约，准确的信道模型可以使网络部署、优化工作更加准确和有效，从而提升无线网络的性能和可靠性[5]。在工业物联网中，许多工业控制系统对网络有着严格的时延和可靠性要求，只有在充分掌握信道和噪声特性之后，才能采取与之相适应的物理层技术并实现合理的系统设计[6]。目前国内外对工业环境下的无线信道和噪声特性研究分析较少，而当前国内工业物联网的发展迫切需要此类的研究。因此，本文对国外相关研究进行回顾与总结，旨在为国内工业环境下的无线信道和噪声特性分析提供一些启发，为我国的工业物联网发展提供理论基础。

2　工业物联网中的无线通信技术

对工业物联网无线通信技术的选择主要取决于其具体应用的场景，因为不同场景对信息传输的功耗、成本、速率、容量等存在差异化的需求。目前，Wi-Fi、蓝牙、ZigBee、RFID（radio frequency identification，射频识别）、UWB（ultra wideband，超带宽）、NFC（near field communication，近场通信）等技术已被广泛应用在短距离无线通信技术中，而 NB-IoT、eMTC 和 LoRa 等新的主流低功耗广域网络技术正在与这些短距离无线通信技术互补，相互配合使用于工业物联网的应用场景中。其中，蓝牙、Wi-Fi 和 ZigBee 技术都使用 2.4 GHz 频段；UWB（ultra wideband，超带宽）是一种无载波通信技术，具有定位精度高、安全性强、抗干扰能力强等特点，主要应用于工厂监控领域；RFID 技术可通过无线电信号识别特定目标，单向读写相关数据，可以提供生产制造控制系统、生产制造执行系统和管理信息系统的服务信息[7]；而 NFC 技术在电子设备之间实现简单和安全的双向交互，应用于物品识别。这些技术各有所长，也各有所短。

NB-IoT 是基于窄带（200 kHz）的蜂窝物联网技术，是专门为低功耗、广覆盖的物联网业务设计的（基于 FDD 模式）。NB-IoT 技术穿墙能力较传统技术有大幅度的提升，且在同一基站情况下，NB-IoT 可以提供现有无线技术 50~100 倍的接入数，一个扇区能够支持 10 万个连接，并且支持低时延敏感度、超低成本、低功耗和优化的网络架构[8]。此外，NB-IoT 构建于蜂窝网络，可直接部署于 GSM（global system for mobile communication，全球移动通信系统）网络、UMTS（universal mobile telecommunications system，通用移动通信系统）网络或 LTE（long term evolution，通用移动通信技术的长期演进）网络，以降低部署成本、实现平滑升级。NB-IoT 继承了 4G 网络的安全能力，支持双向鉴权以及空口严格加密，确保用户数据的安全性和稳定性，有效支撑工业物联网应用。

eMTC 在 LTE 系统的基础上，为低功耗、广覆盖物联网业务拓展了新功能，可在 LTE 系统上实现软件升级。eMTC 支持上下行最大 1 Mbit/s 峰值速率，远远超过传统 GPRS（general packet radio service，通用分组无线服务）、ZigBee 等技术的速率；eMTC 支持连接态移动性，物联网用户可以无缝切换，保障用户体验；基于 TDD（time division duplexing，时分双工）的 eMTC 还可提供低成本的定位技术，在物流跟踪、货物跟踪等场景应用广泛[9]。

LoRa 是一种基于扩频技术的远距离无线传输技术，最早由美国 Semtech 公司采用和推广。LoRa 极大地改善了接收的灵敏度，降低了功耗；LoRa 技术的网关支持多信道多数据速率的并行处理，系统容量大，还可以支持测距和定位。LoRa 技术的特点使其非常适用于要求功耗低、距离远、大量连接及定位跟踪的物联网应用场景[10]。

综上所述，NB-IoT 虽然以低功耗、广域网、低速率、待机时间长而著称，但其带宽只有 200 kHz 左右，不能达到较高速率业务的需求。eMTC 带宽 1.4 MHz，具有良好的移动性和语音功能，可在 LTE 系统上直接升级软件支持。NB-IoT 和 eMTC 适合于面积广阔的公共空间，LoRa 主要使用于非授权频段，受无线覆盖范围限制，适用于一些短距离覆盖和专用网络场景应用，满足很多工业生产生活中对小范围内构建局域网的需求。

3 工业物联网信道测量研究现状与信道特性研究回顾

3.1 工业物联网信道特点

工业无线通信的发展和实施，能够让企业的管理部门和生产现场数据信息进行实时更新，实时掌握工厂中的生产情况，能够更迅速、及时准确地互通信息进行控制和管理。常规工厂环境大致分为 3 类典型工业场景。

（1）精密工业场景

如手机电路板电装车间、家用电器生产线车间等（如图 1（a）所示）。这类场景中，生产都是在机箱内部流水线进行，加工设备在出厂之前需要 3C（China compulsory certification，中国强制认证）认证，对外的电磁辐射等都有限制[11]，因此，工厂内电磁干扰相对较好。而操作工人走动会导致信道的时变性。

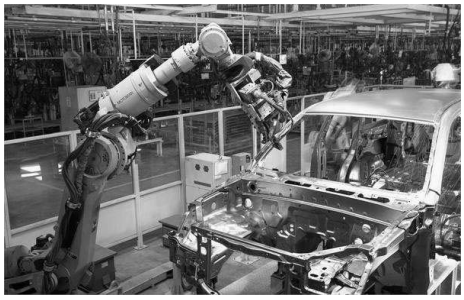
（2）常规工业场景

如汽车加工车间（如图 1（b）所示），这类场景中，变频器、点火系统、稳压器、高压输电线、电子开关等会辐射出大量的电磁噪声[12]。此外，这类传播环境属于时变信道传播环境，原因是厂内有工业机器人、机械臂等自动化设备摇摆工作。

（3）传统工业场景

如钢铁厂车间（如图 1（c）所示）。这类场景中，工业建筑空间相对较大，机器设备尺寸很大，加热炉、摇臂钻床等会产生大量的噪声，同时有大型的输送机、起重机械、装卸机器等工作使这类场景信道具有时变特性。

(a) 精密工业场景（波峰焊车间）



(b) 常规工业场景（汽车加工车间）

(c) 传统工业场景（钢铁厂车间）

图 1　3 种典型工厂场景

通过以上典型工业场景与其他典型场景（市区、办公室、家庭等）无线信道的对比，如果将物联网部署工业环境，工业物联网环境下的无线

信道会表现出如下不同的特点。

（1）不同的工厂存放的材料对信道影响程度不同。例如，造纸厂仓库环境具有高吸收特点，无线电波以直射波传播并且强度快速衰减；钢铁厂环境具有高反射特点，厂内大型金属设备导致电波在信道传播过程中出现衍射和反射特性。

（2）不同工厂的结构差异大，工业厂房比普通的住宅和办公楼的楼层高，普通住宅层高一般为 2.8 m 左右，而标准厂房高度一般是 5~6 m，并且工厂环境比家庭和办公环境恶劣许多，如振动、浮尘等因素都会对信号进行反射和散射，从而产生多径衰落。

（3）出现多普勒频偏现象，工厂内有工人、机器人、卡车、悬挂设备等的随机移动，这会让工厂环境中的无线信道具有时变特性。如汽车厂使用的移动机器人，机器臂的摇摆运动会带来多普勒偏移。此外，由于地面的不平稳，车体运动过程中出现不断的晃动也会产生多普勒随机频偏现象。

3.2　工业物联网信道测量研究现状

无线信号在传播中受环境、地形等因素的影响，使得无线信道的衰落特性变化十分复杂，这将从根本上制约工业无线通信系统的性能[11]，针对工业场景的信道特性研究是工业物联网无线通信系统设计中需要考虑的重要问题。1989 年，参考文献[13]通过数据分析了路径损耗和时延功率谱，在美国印第安纳五座工厂 LOS（line of sight，视距）和 OBS（obstructed line-of-sight，遮挡视线）条件下重复发送一个 10 ns 脉冲信号，测量频率为 1.3 GHz，通过示波器衰减、失真，完成了信道的时域测量。2004 年，参考文献[14]提出了一个在小型焚化炉厂中超宽带（ultra wideband，UWB）信道统计模型，分析了频率为 3.1~10.6 GHz 视距（LOS）和非视距（NLOS）的小尺度衰落和功率时延分布。2005 年，参考文献[15]利用定向天线来研究无线信道，测量频率为 2.4 GHz，在马格德堡基地一个类似于制造厂的环境下分析了平均时延、均方根延迟和相干带宽。2007 年，参考文献[16]讨论了工业环境中 3 个频率的窄带测量，即 0.9 GHz、2.4 GHz 和 5.2 GHz，开发了一种测量程序，并测量了 3 种场景下的路径损耗，得出了影响工业环境信号传播的因素，提出了一种工业路径损耗模型。2009 年，参考文献[17]基于频域分析方法测量了频率为 5.5 GHz 的工厂环境，研究了工厂 LOS 条件下的信道多径分量和时延扩展。2010 年，参考文献[18]在短距离室内无线传感网络环境下，利用矢量网络分析仪测量频率为 5.8 GHz 的信道冲激响应，研究了 LOS 和 NLOS 条件下的均方根时延，并与其他具有特定统计分布的模型进行比较，提出了对室内无线传感器网络和工业物联网应用都有效的短距离衰落模型。2012 年，参考文献[19]研究了两种的重要工业环境对无线电波传播的影响，一种是高吸收环境，另一种是高反射环境，根据无绳电话和工业科学医疗使用频段选择的频率分别为 0.43 GHz、1.89 GHz、2.45 GHz，分析了信道路径损耗和多径分量。参考文献[20]使用矢量网络分析仪和虚拟天线阵列方法测量了频率 0.8～2.7 GHz 的 3 种不同工业环境，分析了信道冲激响应和功率时延分布，并对经典 Saleh-Valenzuela（S-V）模型进行了修改。2016 年，参考文献[21]研究了自动化工厂车间的无线电波传播，使用宽带信道探测仪在 5.85 GHz 载波频率下进行信道测量，对信道时延特性进行统计研究。2017 年，参考文献[22]在 3 个地点进行了 7 个场景测量，研究了频率为 5.8 GHz 的 LOS 和 NLOS 场景，分析了路径损耗指数、RMS（root mean square，均方根）时延扩展、相干带宽和小尺度衰落的幅度分布。

但是到目前为止，国内还没有用于描述工业物联网环境下典型无线信道的分析模型。从过去的研究中可以发现一些原因，首先是传统的信道研究，没有考虑到工厂环境时变特性对信道的影响；其次，工业无线网络中高反射材料对无线信道特性的影响

并未得到足够重视；并且，工厂环境下信道测量往往很难开展，测量数据的缺乏使得很难建立准确可靠的工业物联网无线信道传播模型。

3.3 工业物联网信道特性

3.3.1 大尺度衰落

大尺度衰落是指收发端之间距离的变化引起信号场强的变化。一方面，接收信号强度随着发射机和接收机之间的距离以对数形式变化；另一方面，在收发端之间距离相同的条件下，由于工厂环境中的物体分布不同以及障碍物会导致信号功率损耗，通常表示为发射功率与接收功率之间的比率。

在过去的几年中，已经进行了多种工业环境下的室内信道测量。表 1 总结了各种典型工业场景大尺度衰落参数。

表 1 各种典型工业场景大尺度衰落参数

测量方法	场景		频率/GHz	路径损耗指数	阴影衰落标准差	适用范围/m
窄带信道	木厂和金属加工厂[23]	LOS	0.9	2.25	5.65	15~140
		OBS[1]	0.9	1.94	4.97	
		OBS[2]	0.9	2.16	5.16	
		LOS	2.4	1.72	4.73	15~140
		OBS[1]	2.4	1.52	4.61	
		OBS[2]	2.4	1.69	6.62	
		LOS	5.2	1.25	4.32	15~140
		OBS[1]	5.2	0.68	3.87	
		OBS[2]	5.2	1.35	3.16	
	工厂[24]	LOS	0.9	2.3	5.7	15~140
		OBS[1]	0.9	2.0	5.0	
		OBS[2]	0.9	2.2	5.2	
宽带信道	混凝土厂[25]	LOS	0.315	2.70	12.65	-
			0.434	2.62	10.60	
			0.87	2.96	11.85	
			0.915	2.98	10.76	
	金属零件的焊接厂[26]	LOS	0.315	2.70	12.65	-
			0.434	2.62	10.60	
			0.87	2.96	11.85	
			0.915	2.98	10.76	
	印刷厂[25]	NLOS	0.315	4.84	14.71	-
			0.434	5.04	14.98	
			0.87	4.39	14.16	
			0.915	4.34	13.93	
	石油钻井平台[26]	LOS	2.4	1.40	1.82	0~10
		NLOS[1]	2.4	2.06	2.17	
		NLOS[2]	2.4	1.17	1.22	
		LOS	5.8	1.76	1.83	0~10
		NLOS[1]	5.8	2.44	2.45	
		NLOS[2]	5.8	1.41	1.31	
	工厂[16]	NLOS	3.1	1.1	1.1	1~30
	工厂[13]	混合	2.2	2.2	7.9	5~100

表 1 总结了各种工业测量中的路径损耗相关参数，其中 OBS[1] 与 OBS[2] 表示遮挡程度不同。在环境、频率和链路配置方面，观察到参数有很大的不同。对于许多环境，路径损耗指数范围在 1~3。另外，随着频率的增加，反射体增加，会导致路径损耗指数降低。然而，不同的工业环境，路径损耗指数和频率之间没有明确的关系。因为路径损耗不仅取决于路径长度，而且还与工业建筑材料类型（金属、木材和混凝土等）、散射体的大小和密度有关。

3.3.2　时间色散

时间色散主要是因为多径传播造成信号时间扩散现象。与其他典型室内环境相比，时间色散在工业环境下无线信道会有很大的不同，时间色散受发射机、接收机和工厂物理环境等因素的影响。PDP（power delay profile，功率时延分布）可以用于定量地描述时间色散。在参考文献[27]中，可以得出电波在传播过程中受建筑物大小、密度、结构、室内地板布局和室内装饰位置等影响。在参考文献[28]中在 LOS 条件下进行测量，从测量结果得到随着收发端天线间距越来越远，时延也随之增加。由以上研究，可以发现，工厂环境无线信道时间色散取决于环境中散射体的大小、类型、密度和分布，并且多径分量到达时间与接收天线间距有关。在参考文献[13]中，对印第安纳州 5 家大型工厂进行测量，每个工厂具有不同特点，所有工厂均方根时延平均值在 LOS 条件下大于 96 ns；在 NLOS 条件下大于 105 ns。数据表明，在工厂建筑物中的无线传播可以通过混合的几何或者统计模型来进行适当的描述，要考虑到墙壁和天花板的镜面反射以及来自仓库物品和设备产生的随机散射。因为在工厂环境中，建筑物的年龄、物品分布、墙壁位置和天花板高度都是影响均方根时延的关键因素。在参考文献[29]中，随着接收天线间距增加，均方根时延扩展增加。在参考文献[15]中图 3 对比了 4 种典型场景下测量场景的信道冲激响应。从平均时延、均方根时延和相干带宽得到的数据结果发现，受干扰的因素不完全依赖于 LOS 条件，与金属阻碍物的分布有关。在静态模式下，RMS 时延扩展超过 72 ns，这表明了金属设备的反射是主要影响因素；当接收天线间距较短时，均方根时延扩展几乎是恒定的。当发射机天线与接收机天线逐渐移开时，反射信号幅度变大，均方根时延增加。

由以上均方根时延测量研究可以发现工厂内设施、天线间距具有不同的反射水平。因此，从一个环境获得的测量结果不适用其他环境，这会使无线技术的可靠性复杂化。

3.3.3　时变特性

如今的工业中，工厂自动化随处可见，厂内人员移动、机器人运动和小车在行驶中摆动等因素使工业环境中的无线信道具有时变特性，会出现多普勒频移现象。一方面，工厂内人员移动、机器人运动等会引起多普勒频偏。在参考文献[30]中，将人体建模为垂直定向的圆柱体，通过将来自室内环境的地板、天花板和墙壁的反射以及来自移动人体散射的多径分量进行参数化，符合 Rice（莱斯）分布并且信号随着工作人员的移动而变化。参考文献[33]研究了工业中有机器人工作的无线信道的时变特性，测量的接收天线安装在做周期圆周运动的机器人手臂上。该机器人手臂重复运动的周期为 1.5 s，运动线速度为 2 m/s，中心频率为 2.44 GHz。测量时间为 10 s，从多普勒频移的测量结果发现，在前 6 s 变化快速，在后 4 s 变化缓慢，然后机器人停止工作，由此得到工业中机器人手臂在做周期性圆周运动时会使信道具有时变特性。另一方面，课题组研究了某汽车厂 LOS 场景下运输小车的信道时变特性，研究发现运输小车的多普勒频移并不是理论上的纯多普勒值，而是在理论值的基础上还存在一定的随机频偏。图 2 是某汽车厂测量场景，沿箭头从左向右匀速推动运输小车，在移

动运输小车的过程中，工厂地面不平，使得小车有不规则摆动。

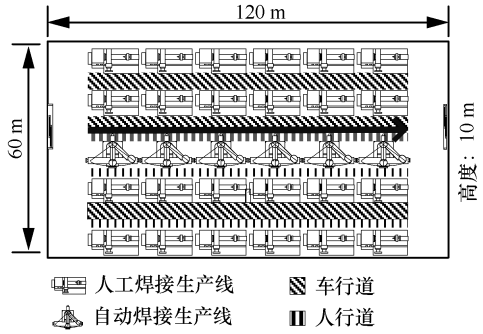


图 2　汽车厂实测场景

根据测量频率为 5.8 GHz 的实测数据将运输小车移动对信道的影响进行了分析，如图 3 所示为瞬间多普勒功率谱，图 3(a)是 LOS 条件下的仿真数据，图 3(b)为 LOS 条件下的实测数据。从测量结果得到运输小车不规则的摆动会导致信号产生有波动频差，并根据实测结果对工业环境下运输小车的多普勒频偏提出了一种数学模型，该模型的多普勒频移符合理论的多普勒频移加具有一定高斯分布的随机频偏值，工厂环境下运输小车多普勒频偏满足：

$$f_{\mathrm{d}}'(t) = f_{\mathrm{d}}(t) + \Delta f \tag{1}$$

其中，$f_{\mathrm{d}}(t)$是一定速度下的理论多普勒频移，即 $f_{\mathrm{d}}(t) = \dfrac{v(t)\cdot\cos\alpha(t)}{\lambda}$，其中 Δf 是均值为 0、方差为 σ_2 的高斯分布变量。这些研究表明，工厂内工作人员的移动导致显著的信号变化；如果链路被卡车和大型机器人穿过，则会使信号出现更多的变化；在工厂工作过程中，有其他因素使得小车或其他设备有不规则的摆动，将会导致信号产生随机波动频差。

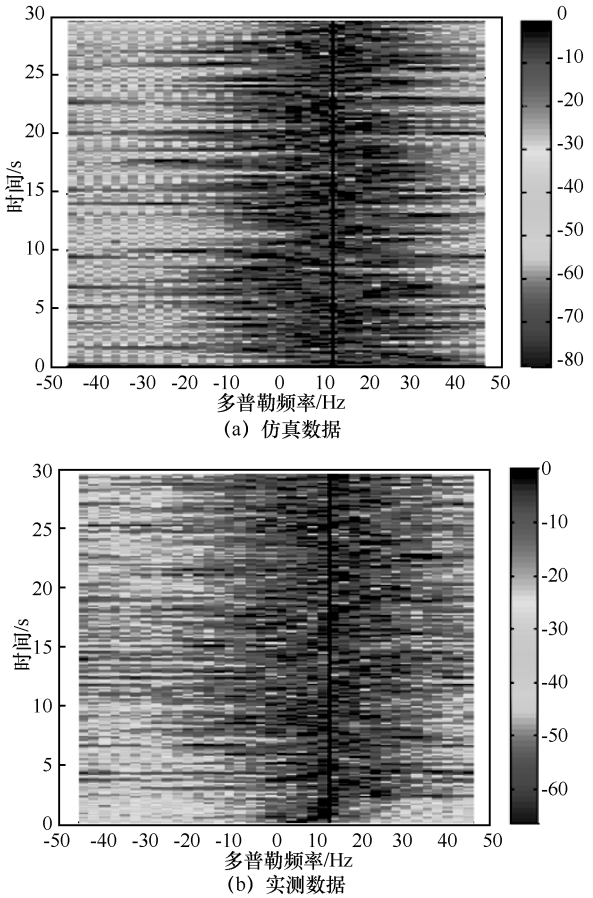


图 3　瞬间多普勒功率谱

4　工业物联网电磁噪声特点与噪声特性研究现状

4.1　工业物联网电磁噪声特点

噪声对无线通信系统的影响十分显著，分析噪声的特点对工业无线网络的稳定性设计十分重要。噪声工业环境下与其他典型环境（市区、办公室、家庭等）也有所不同，其特点主要分为以下几个方面。

（1）工厂噪声的种类较多，对无线信号产生影响的噪声主要有机械性噪声和电磁性噪声。机械性噪声由机械撞击、摩擦、转动而产生，如破碎机、球磨机、电锯和机床等发出的噪声；电磁性噪声由于磁场脉动、电源频率脉动引起电器部件震动而产生，如发电机、变压器、继电器产生的噪声。

（2）工厂噪声的分布频率范围较大，且各种类型的噪声总是同时存在于工厂内各个随机的位置。比如焊接产生的噪声频率可以达到数百 MHz，火花放电产生的电磁噪声辐射分布在几百 MHz 到 GHz，用于控制机械设备的计算机产生的电磁

辐射频率分布在几十 Hz 到 3 GHz。

（3）工厂噪声不仅有高斯白噪声影响，还有突发的脉冲噪声等，且工厂环境的噪声在强度、频率范围、带宽、功率谱形状、生灭特征等方面与传统的高斯白噪声存在很大的差异；在工厂正常工作时，由于机械设备的间歇性工作，其辐射出的电磁噪声也会呈现出一定的时变特性。

4.2　工业物联网电磁噪声研究现状

在参考文献[31]中提出了工业无线网络的宽带信道模型，该模型考虑了恶劣工厂环境中噪声的影响，采用一阶两态马尔可夫过程描述工业环境中典型突发脉冲噪声的特性。脉冲噪声的幅值与噪声功率比有关，当噪声功率变大时，脉冲噪声的幅值增大。从测试结果中发现，如果没有考虑噪声对信道的影响，工厂环境的噪声会极大地降低无线通信系统的性能。在参考文献[32]中测量了 100 MHz~6 GHz 频段下变电站的电磁噪声环境，给出了脉冲率、脉冲幅度、脉冲持续时间、脉冲发生时间等统计分析，构建了基于统计和频谱特性的脉冲噪声模型，用于评估变电站无线设备的部署相关情况。在参考文献[33]中提出了用 APD（amplitude probability distribution，幅度概率分布）统计方式评估电磁噪声对通信系统的影响，将幅度概率分布定义为干扰强度超过某个电平的时间概率，计算式为：

$$\mathrm{APD}(x_0)=\Pr[X>x_0]=1-F(x_0) \qquad (2)$$

其中，$F(x_0)$ 是 X 的 CDF（cumulative distribution function，累积分布函数）。通过幅度概率分布统计参量的测量结果，可以获得噪声电平的平均值和有效值，能够真实地反映噪声的特性，评估噪声对不同类型的通信系统的影响。在参考文献[30]中，用 APD 测量的环境是典型钢铁厂，其中心频率分别为 439 MHz、440 MHz、570 MHz 及 2 450 MHz，其中，干扰在 439 MHz 时最明显，因为有一辆汽车、一辆运输机器人和一台起重机同时工作。在造纸厂电动机将大块木头粉碎的过程中，产生的噪声使得 DECT（digital enhanced cordless telecommunication，室内无绳电话）系统突然出现问题无法使用。本文对某汽车厂焊接设备的电磁噪声展开了研究，采用对数周期天线分别测量了手工点焊机与焊接机器人附近的噪声信号，得到了噪声功率、噪声带宽、噪声之间频率间隔等相关参数。图 4（a）和图 4（b）分别为手工点焊机与焊接机器人附近的噪声信号，从测量结果中可以看出，手工点焊机与焊接机器人工作时产生的噪声分布大体上相同，但并不完全吻合。两种焊接设备的噪声信号主要分布在 300~900 MHz 频段，其功率主要集中在-110~97 dBm，带宽在 4~20 kHz 范围相对较窄。这主要是由于两种机械设备采用的焊头相同，故噪声情况相近，但工作时焊头的高度、焊机工作方式略有差异，导致噪声略有不同，但两者总体上大致相似。由这些研究可以看出，不同工厂噪声差异很大，对于噪声测量应全面考虑工业物联网所应用的频段。若影响无线通信系统的噪声只用加性高斯白噪声（additive white Gaussian noise）表示，在恶劣的工厂环境中存在的脉冲噪声会显著降低无线系统的可靠性和有效性。

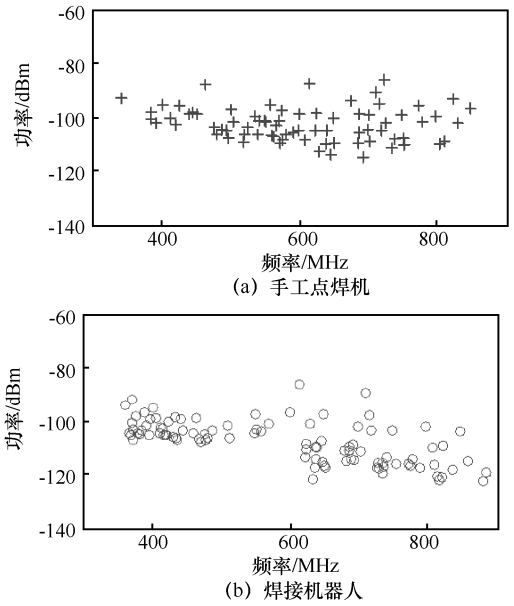


图 4　手工点焊机与焊接机器人附近的噪声信号

5 结束语

工业物联网时代已经开始，想要实现自动化工厂，需要加速完善通信技术来改进生产流程和管理系统。本文首先对工业物联网中使用的无线通信技术进行总结和对比，归纳了工业物联网环境下的信道和电磁噪声特点，回顾了工业物联网信道和噪声测量以及研究的已有成果。本文讨论了不同工业环境条件下的路径损耗，分析了信道时间色散、时变特性以及噪声特性，并强调了影响它们的因素。研究发现只有充分了解无线信道和噪声特性以及不同无线技术适用的工业场景，才能使工业生产中不可预测的干扰风险降至最低。工业物联网的发展离不开无线通信技术的支持，未来迫切需要对不同工厂环境、不同频率和链路配置的信道和噪声测量以及对其特性展开分析，以进一步验证现有的模型或开发新的模型，适应工业物联网的迅猛发展。

参考文献:

[1] 杜玉河. 《工业物联网白皮书》正式发布[J]. 起重运输机械, 2017(10): 34-35.
DU Y H. The industrial internet of things white paper was officially released [J]. Lifting and Transport Machinery, 2017(10): 34-35.

[2] 工业和信息化部. 物联网的十三五规划(2016-2020)年[R]. 2016.
Ministry of Industry and Information Technology. IoT 13th five-year plan (2016-2020) years[R]. 2016.

[3] 陈山枝. 发展 5G 的分析与建议[J]. 电信科学, 2016, 32(7): 1-10.
CHEN S Z. Analysis and suggestion of future 5G directions [J]. Telecommunications Science, 2016, 32(7): 1-10.

[4] 宫诗寻, 陶小峰. 5G 大规模机器类通信中的传输技术[J]. 中兴通讯技术, 2017, 23(3): 20-23.
GONG S X, TAO X F. Transmission technologies in massive machine type communication for 5G[J]. ZTE Technology Journal, 2017, 23(3): 20-23.

[5] 刘留, 陶成, 余立, 等. 高速铁路无线信道测量与信道模型探讨[J]. 电信科学, 2011, 27(5): 54-60.
LIU L, TAO C, YU L, et al. Discussion on the channel measurement and channel model under high speed railway environment [J]. Telecommunications Science, 2011, 27(5): 54-60.

[6] 赖春媛, 闫文卿, 亓晋. 基于云雾融合的工业物联网能源管理架构[J]. 电信科学, 2017, 33(10): 2-9.
LAI C Y, YAN W Q, QI J. An energy management framework based on fog-cloud combining for industrial internet of things [J]. Telecommunications Science, 2017, 33(10): 2-9.

[7] 常洁, 王艺, 李洁, 等. 工业通信网络现有架构的梳理总结和未来运营商的发展策略[J]. 电信科学, 2017, 33(11): 123-133.
CHANG J, WANG Y, LI J, et al. Summary of existing framework in industrial communication networks and future development strategies for communication operators [J]. Telecommunications Science, 2017, 33(11): 123-133.

[8] SKYLAB. 物联网时代的黑马——NB-IoT[R]. 2017.
SKYLAB. The black horse in the age of the internet of things——NB-IoT[R]. 2017.

[9] 邢宇龙, 张力方, 胡云. 移动蜂窝物联网演进方案研究[J]. 邮电设计技术, 2016(11): 87-92.
XING Y L, ZHANG L F, HU Y. The evolution research of mobile cellular communication technology for IoT[J]. Designing Techniques of Posts and Telecommunications, 2016(11): 87-92.

[10] 严小强, 李星. 基于LoRa功耗及其响应速度设计研究[J]. 电子质量, 2017(9): 9-12.
YAN X Q, LI X. Research on power consumption and response speed design based on LoRa[J]. Electronics Quality, 2017(9): 9-12.

[11] 韩磊. 移动用户终端的电磁辐射发射测试软件开发[D]. 北京: 北京交通大学, 2011.
HAN L. The development of electromagnetic radiation emission test software for mobile user terminal[D]. Beijing: Beijing Jiaotong University, 2011.

[12] CHEFFENA M. Propagation channel characteristics of industrial wireless sensor networks [wireless corner][J]. IEEE Antennas & Propagation Magazine, 2016, 58(1): 66-73.

[13] RAPPAPORT T S. Characterization of UHF multipath radio channels in factory buildings[J]. IEEE Transactions on Antennas & Propagation, 1989, 37(8): 1058-1069.

[14] KAREDAL J, WYNE S, ALMERS P, et al. Statistical analysis of the UWB channel in an industrial environment[C]//Vehicular Technology Conference (VTC2004-Fall), Sept 26-29, 2004, Los Angeles, CA, USA. Piscataway: IEEE Press, 2004: 81-85.

[15] MIAOUDAKIS A, LEKKAS A, KALIVAS G, et al. Radio channel characterization in industrial environments and spread spectrum modem performance[C]//2005 IEEE Conference on Emerging Technologies and Factory Automation (ETFA 2005), Sept 19-22, 2005, Catania, Italy. Piscataway: IEEE Press, 2005: 93.

[16] KAREDAL J, WYNE S, ALMERS P, et al. A measurement-based statistical model for industrial ultra-wideband channels[J]. IEEE Transactions on Wireless Communications, 2007, 6(8): 3028-3037.

[17] KOZLOWSKI S, SZUMNY R, KUREK K, et al. Statistical modelling of a wideband propagation channel in the factory environment[C]//2009 European Conference on Wireless Technology, Sept 28-29, 2009, Rome, Italy. Piscataway: IEEE Press, 2009: 190-193.

[18] WANG Y, LU W, ZHU H. Experimental study on indoor channel model for wireless sensor networks and Internet of Things[C]//2010 IEEE International Conference on Communication Technology, Nov 11-14, 2010, Nanjing, China. Piscataway: IEEE Press, 2010: 624-627.

[19] FERRER-COLL J, ÄNGSKOG P, CHILO J, et al. Characterisation of highly absorbent and highly reflective radio wave propagation environments in industrial applications[J]. Communications IET, 2012, 6(15): 2404-2412.

[20] AI Y, CHEFFENA M, LI Q. Power delay profile analysis and modeling of industrial indoor channels[C]//2015 European Conference on Antennas and Propagation, April 13-17, 2015, Lisbon, Portugal. [S.l.: s.n.], 2015: 1-5.

[21] HOLFELD B, WIERUCH D, RASCHKOWSKI L, et al. Radio channel characterization at 5.85 GHz for wireless M2M communication of industrial robots[C]//2016 IEEE WCNC, May 23-27, 2016, Kuala Lumpur, Malaysia. Piscataway: IEEE Press, 2016.

[22] CROONENBROECK R, UNDERBERG L, WULF A, et al. Measurements for the development of an enhanced model for wireless channels in industrial environments[C]// IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, Oct 9-11, 2017, Rome, Italy. Piscataway: IEEE Press, 2017: 1-8.

[23] TANGHE E, JOSEPH W, VERLOOCK L, et al. The industrial indoor channel: large-scale and temporal fading at 900, 2400, and 5200 MHz[J]. IEEE Transactions on Wireless Communications, 2008, 7(7): 2740-2751.

[24] TANGHE E, JOSEPH W, MARTENS L, et al. Large-scale fading in industrial environments at wireless communication frequencies[C]// 2007 Antennas and Propagation Society International Symposium, June 10-15, 2007, Honolulu, Hawaii, USA. Piscataway: IEEE Press, 2007: 3001-3004.

[25] ŞEYMA TÜTÜNCÜ, KARA A. UHF propagation measurements in heavy industry[C]//2016 Signal Processing and Communication Application Conference, May 16-19, 2016, Zonguldak, Turkey. Piscataway: IEEE Press, 2016.

[26] LUO S, POLU N, CHEN Z, et al. RF channel modeling of a WSN testbed for industrial environment[C]//2011 Radio & Wireless Symposium, Jan 1, 2011, Phoenix, USA. Piscataway: IEEE Press, 2011: 375-378.

[27] WITTMANN M, MARTI J, KURNER T. Impact of the power delay profile shape on the bit error rate in mobile radio systems[J]. IEEE Transactions on Vehicular Technology, 2002, 46(2): 329-339.

[28] KHAN H N, GROSINGER J, GÖRTSCHACHER L, et al. Statistical analysis of the power delay profile of a SIMO UHF backscatter RFID channel in an engine test bed[C]// 2017 Antennas & Propagation Conference, June 9-14, 2017, San Diego, CA, USA. Piscataway: IEEE Press, 2017: 1-5.

[29] AI Y, CHEFFENA M, LI Q. Radio frequency measurements and capacity analysis for industrial indoor environments[C]// 2015 European Conference on Antennas and Propagation, Apr 12-17, 2015, Lisbon, Portugal. Piscataway: IEEE Press, 2015.

[30] KAREDAL J, WYNE S, ALMERS P, et al. UWB channel measurements in an industrial environment[C]//2004 Global Telecommunications Conference (GLOBECOM'04), November 29-December 3, 2004, Dallas, Texas, USA. Piscataway: IEEE Press, 2004: 3511-3516.

[31] CHEFFENA M. Industrial wireless sensor networks: channel modeling and performance evaluation[J]. EURASIP Journal on Wireless Communications & Networking, 2012(1): 1-8.

[32] SHAN Q, BHATTI S, GLOVER I A, et al. Characteristics of impulsive noise in electricity substations[C]// 2009 Signal Processing Conference, Aug 24-28, 2009, Glasgow, Scotland, UK. Piscataway: IEEE Press, 2009: 2136-2140.

[33] 杨飞, 阚润田, 沙斐. 无线电骚扰的统计测量方法研究[J]. 电子测量与仪器学报, 2009, 23(1): 22-26.
YANG F, KAN R T, SHA F. Research on statistical measurement of radio noise and electromagnetic disturbance[J]. Journal of Electronic Measurement and Instrument, 2009, 23(1): 22-26.

[作者简介]

张克（1994−），女，北京交通大学硕士生，主要研究方向为宽带无线通信。

刘留（1981−），男，博士，北京交通大学电子信息工程学院教授、博士生导师，主要研究方向为高铁无线信道测量与建模、高铁宽带接入物理层关键技术等。

袁泽（1994−），男，北京交通大学硕士生，主要研究方向为宽带无线通信。

张琨（1993−），男，北京交通大学硕士生，主要研究方向为宽带无线通信。

张建华（1976−），女，北京邮电大学信息与通信工程学院教授、博士生导师，主要方向为宽带移动通信系统新理论及技术等。

张志军（1986−），男，北京航天测控技术有限公司高级工程师，主要研究方向为测控技术、装备自动化测试等。

研究与开发

基于迁移学习的室内动态环境定位算法

刘参，尚俊娜，李蕊江，岳克强
（杭州电子科技大学，浙江 杭州 310018）

摘　要：传统室内指纹定位系统的精度受指纹库中参考位置节点的密度和室内环境特征等多方面因素的制约。室内环境动态变化时 RSS 波动较大，通常不满足同分布的假设条件，故传统指纹定位方法难以满足高精度需求。针对室内环境动态变化导致传统算法无法精准定位问题，设计并实现了一种基于室内指纹库的迁移学习动态环境定位算法，该算法采用迁移学习的思想把不同分布的数据集嵌入对齐到潜在特征空间中，从而有效缓解了环境动态变化对系统造成的不利影响。本文算法实验数据均来自于真实的环境，通过仿真得到该算法的平均定位误差是 1.23 m。

关键词：动态定位；迁移学习；室内环境特征；广义延拓插值；RSS 指纹库；低工作量
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018170

Indoor dynamic environment localization algorithm based on transfer learning

LIU Can, SHANG Junna, LI Ruijiang, YUE Keqiang
Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: The accuracy of the traditional indoor fingerprint localization system is limited by many factors, such as the density of the reference location node in the fingerprint database and the characteristics of the indoor environment. When the indoor environment changes dynamically, the RSS fluctuates, and usually does not meet the assumption of the same distribution. Therefore, it was difficult to obtain high-precision requirements for conventional fingerprint positioning method. Aiming at the problem that the traditional algorithm couldn't locate accurately, an algorithm based on the indoor fingerprint database was designed and implemented. The algorithm adopted the idea of migration learning to embed different data sets into the latent feature space, and the adverse effects of environmental changes on the system were mitigated. The simulation results show that the average positioning error of this algorithm is 1.23 m.

Key words: dynamic localization, transfer learning, indoor environmental characteristic, generalized extended interpolation, received signal strength fingerprint, reduced calibration effort

收稿日期：2018−01−09；**修回日期**：2018−05−04
通信作者：尚俊娜，shangjn@hdu.edu.cn
基金项目：国家自然科学基金资助项目（No.11603041）
Foundation Item: The National Natural Science Foundation of China (No.11603041)

1　引言

随着无线传感器网络（wireless sensor network，WSN）技术的发展，室内定位技术受到越来越多的关注，人们对室内无线定位的需求日益增大，如仓库、超市、地下停车场、机场大厅、监狱等，常常需要知道物品或行人的位置[1-2]。基于接收信号强度（received signal strength，RSS）的定位系统不用增加额外的硬件设备，同时具有低成本、低功耗等优势，成为国内外研究的热点[3]。基于 RSS 的定位系统可分为信号传播模型[4]和无线电指纹[5]两种，由于室内环境复杂多变，特别是行人、桌椅等家具移动使得 RSS 具有较大的波动性[6]，采用滤波的方法可以减缓这种波动性，但是该方法仍难以解决室内物体变化后的动态环境定位问题。故此，本文采用基于 RSS 指纹库的动态环境定位方法，避免了信号传播模型不确定性造成的定位精度下降的问题。

基于指纹库的定位系统通常分为在线训练阶段和离线定位阶段，系统的定位精度取决于训练阶段指纹数据库的质量[7]，包括指纹库中参考节点的采样密度和指纹信息的准确性。随着参考节点数目的增加，指纹信息更能反映出实际室内环境中 RSS 量的变化趋势，故系统的定位精度更高，但在指纹库创建时需要消费巨大的人力和财力资源[5]。基于指纹库的定位系统，通常假设训练数据和测试数据服从相同的分布[8]，在实际室内复杂环境中，障碍物的遮挡、信号的干扰以及移动物体的运动等，导致原始 RSS 指纹库中的数据失效，此时用已失效的指纹数据进行实时动态定位，定位精度普遍较低。故此，低成本、高精度的动态定位系统受到越来越多学者的关注。参考文献[9]提出对路径衰减因子进行动态修复的质心定位算法，参考文献[10]采用基于线性内插法的室内指纹定位算法，可以减缓室内复杂环境对定位性能的影响，但是当动态环境发生较大变化后，系统的定位精度较低；参考文献[11]提出针对动态环境的自适应定位方法，但是该算法需要在变化后的环境中采集固定位置处标签节点的信息，会带来额外的成本开支，无法得到普遍应用；参考文献[12]提出了基于隐性马尔可夫模型的 Ma.HMM 定位算法，但此方法需要额外的辅助数据更新指纹库且计算复杂度较高；参考文献[13]提出利用用户状态的方法动态更新环境变化后的指纹库，但是需要粘贴二维码标志，整个系统后期维护较困难，无法在大型场所广泛应用。

为了有效缓解室内环境动态变化对定位系统精度造成的不利影响，并且进一步减小工作量，本文提出了一种基于迁移学习的室内动态环境定位算法。在该算法中，首先进行室内 RSS 指纹库的创建，为了节约成本，本文设计了一种可以明显减少工作量同时不损失指纹库质量的广义延拓插值算法；在进行动态环境定位时，利用 RSS 信息在潜在特征空间的空间关联性[14]，本文设计了一种迁移学习动态环境定位算法，通过在环境变化以后的空间中随机采样，找到潜在的特征空间根据领域自适应算法进行定位，进而消除室内环境变化对系统定位性能的影响。最后通过大量的试验仿真，证明本文所提的算法相比传统 *k*NN 算法[15]和 Ma.HMM 算法[12]，在定位精度和系统稳定性方面都有较大的改善。故此，本文算法可实现低成本、高精度的室内动态环境定位。

2　相关工作

传统基于指纹的机器学习定位算法假设训练数据和测试数据服从相同的分布，此时训练学习到的 RSS 映射模型直接应用在不同时间段内进行位置的估计。然而在实际室内复杂动态环境中，障碍物的遮挡、无线电信号的干扰以及移动物体的不可预测运动等，使得 RSS 具有明显的不确定性，因此在多数情况下并不满足上述的同分布假设[16]。当动态环境变化时，RSS 映射模型也会发

生变化，因此基于指纹的定位系统无法使用预先训练得到的模型进行高精度定位。

2.1 室内 RSS 指纹分布特性

经典的 RSS 对数衰减模型[17]如式（1）所示：

$$\mathrm{RSS} = R_{\mathrm{o}} - 10 \cdot n \cdot \lg d + X_{\sigma} \tag{1}$$

其中，R_{o} 表示接收端与锚节点相距 1 m 时的 RSS 值，n 表示路径衰减因子，d 表示测试节点与锚节点之间的距离，X_{σ} 表示环境噪声，通常服从 $\mathcal{N}\left(0,\sigma^2\right)$ 的高斯分布，σ 表示环境噪声的大小，环境噪声越大，信号传播路径损耗越严重。A 和 n 都是经验值，与具体的室内环境有关，参考文献[17]研究了不同环境下 RSS 衰减模型参数的取值范围。

根据经验可知，在相对距离接近的参考位置处，接收到的 RSS 信号彼此相似；虽然在不同时间段内 RSS 具有明显的波动性，但在同一时间范围内，这种波动很小。图 1 是在实际室内动态环境下，同一位置不同时间段内的 RSS 分布直方图。

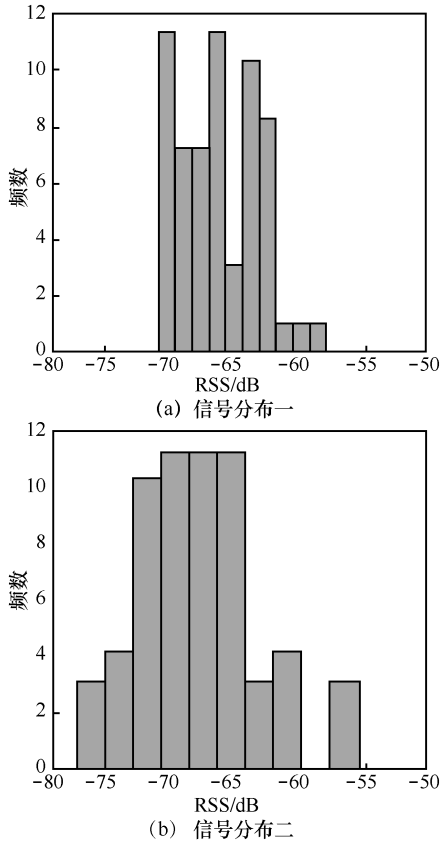


图 1 同一位置不同时间段内的 RSS 分布直方图

从图 1 可知，在固定位置处，不同时间段内的 RSS 分布是不同的，因此基于指纹的室内定位方法采用训练学习得到的 RSS 映射模型进行不同分布情况下的位置估计，会得到较大的定位误差，无法实现动态环境室内高精度定位功能[18-19]。此时需要采用迁移学习的思想，把前一时间段内训练得到的 RSS 映射模型根据流形对齐理论迁移学习到不同数据分布的时间段内，以此来实现自适应动态定位功能[20]。

2.2 迁移学习动态定位实例

$$\text{RSS 矩阵 } S_{\mathrm{RSS}}^{\mathrm{A}} = \begin{bmatrix} \mathrm{rss}_1^1 & \mathrm{rss}_1^2 & \cdots & \mathrm{rss}_1^{N_{\mathrm{A}}} \\ \mathrm{rss}_2^1 & \mathrm{rss}_2^2 & \cdots & \mathrm{rss}_2^{N_{\mathrm{A}}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{rss}_{M_{\mathrm{A}}}^1 & \mathrm{rss}_{M_{\mathrm{A}}}^2 & \cdots & \mathrm{rss}_{M_{\mathrm{A}}}^{N_{\mathrm{A}}} \end{bmatrix} \in \mathbb{R}^{M_{\mathrm{A}} \times N_{\mathrm{A}}}$$

作为训练数据集，其中，rss_i^j $(i = 1,2,\cdots,M_{\mathrm{A}}$，$j = 1,2,\cdots,N_{\mathrm{A}})$ 表示在第 i 个参考位置处接收到第 j 个锚节点（AP）的信号强度值，M_{A} 表示锚节点的个数，N_{A} 表示参考位置的个数。

从 $S_{\mathrm{RSS}}^{\mathrm{A}}$ 矩阵可以看出，在某一参考位置处，存在唯一的接收信号强度向量，可以认为信号强度空间与真实的物理位置空间存在某种映射关系。同时可以发现 $S_{\mathrm{RSS}}^{\mathrm{A}}$ 矩阵隐含的 3 个重要特征[21]：每一列代表在某参考位置处接收的 RSS 指纹信息，如果某两列 RSS 指纹信息比较相似，表示该数据来自两个离得较近的位置；每一行代表某个锚节点在所有参考位置处的 AP 指纹信息，如果某两行数据彼此相似，则表示这两个 AP 的位置离得较近；矩阵中的每个元素 rss_i^j 表示第 i 个参考位置处接收到第 j 个 AP 的信号强度，如果信号强度具有最大值，则表示参考位置 i 离第 j 个 AP 的位置最近。

将在不同时段内随机采样的数据作为测试数据集 $S_{\mathrm{RSS}}^{\mathrm{B}} \in i^{\,M_{\mathrm{B}} \times N_{\mathrm{B}}}$，其中，$M_{\mathrm{B}}$ 表示锚节点的个数，N_{B} 表示随机采样的采样数据数。由于不同时间段内的数据服从不同的分布，故此传统机器学习算法求解到的 RSS 映射模型不再适用于测试数据

$S_{\mathrm{RSS}}^{\mathrm{B}}$，然而在同一室内环境中，锚节点的位置通常是固定的，可认为这些锚节点在同一室内场景中属于共享锚节点（shared anchor point，SAP），上述 $S_{\mathrm{RSS}}^{\mathrm{A}}$ 矩阵隐含的 3 个重要特征同样适用于测试数据 $S_{\mathrm{RSS}}^{\mathrm{B}}$。

室内动态环境中 RSS 信号具有不同的分布特性，但是在邻近的物理位置处，不同分布下的 RSS 信号是彼此相似的，故可以把训练数据集 $S_{\mathrm{RSS}}^{\mathrm{A}}$ 与测试数据 $S_{\mathrm{RSS}}^{\mathrm{B}}$ 在某潜在的低维空间中进行流形对齐，即认为这些不同分布的数据在某个潜在空间中具有相似的特性[18,22]，例如这些数据都与真实的物理位置相对应。以具有 2 个 SAP 的二维信号空间为例，如图 2 所示，在信号空间 A 和信号空间 B 内信号分布是不同的，但是在潜在的二维物理位置空间里面存在映射对齐关系[22]，图 2 中含有位置信息的向量 R_2 来自信号空间 A，未含有位置信息的向量 R_2' 来自信号空间 B，向量 R_2 和 R_2' 来自不同的信号空间，却都来自同一个物理位置空间 L_2，由于信号空间 B 内未含有物理位置坐标的标签信息，故无法把向量 R_2' 与物理位置坐标 L_2 进行映射对齐。由于同样的原因，无法把向量 R_3'、R_4'、R_5' 与真实的位置坐标 L_3、L_4、L_5 进行对齐。由上述信号矩阵隐含的 3 个重要特征可知，标签数据向量 R_1 在信号空间 A 中接收到来自 AP_1 的信号强度具有最大值，故 R_1 的真实位置靠近 AP_1，而在信号空间 B 中，向量 R_1' 接收到来自 SAP_1 的信号强度具有最大值，R_1' 的位置更靠近 AP_1，可认为向量 R_1 和 R_1' 存在固有的对应关系，称为 RSS 相关对 $\{R_1, R_1'\}$。同理，可以找到来自不同信号空间的更多 RSS 相关对 $\{R_i, R_j'\}$，通过这种对应关系，把 R_i 中含有位置坐标的标签信息传递给未含有位置坐标的 R_j'，实现 R_j' 的位置坐标的标定工作。

通过使用相同类内部与不同类之间的领域关系，把标签数据的位置坐标信息传递到未标签数据的对应物理位置处，例如把信号空间 A 的标签向量

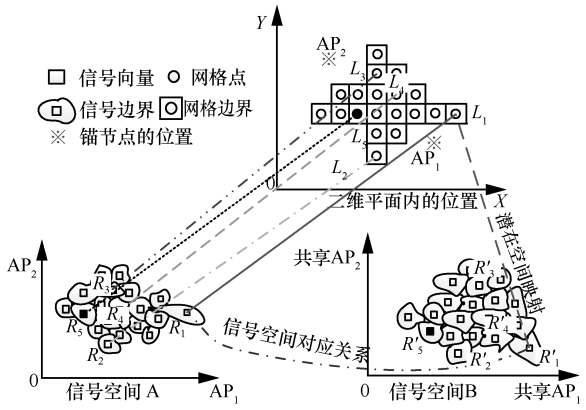


图 2　位置空间与两个不同分布数据空间的相关性

R_1、R_2、R_3、R_4、R_5 的位置信息传递给信号空间 B 的未标签向量 R_1'、R_2'、R_3'、R_4'、R_5'，将上述未标签向量对齐到真实位置坐标 L_1、L_2、L_3、L_4、L_5 上，以此来实现迁移学习动态定位。

3　迁移学习动态定位算法

图 3 是本文所提出算法示意图，为了实现上述迁移定位功能，需要以下几个步骤。

步骤 1　数据采集——广义延拓插值算法

在原始环境中进行轻工作量、低采样密度的指纹库采集，利用广义延拓插值算法得到高质量的指纹库，作为含有位置信息的训练数据 $X_{\mathrm{tra}}^{\mathrm{A}}$，$X_{\mathrm{tra}}^{\mathrm{A}}=\left\{\left(s_i^{\mathrm{A}},\ell_i^{\mathrm{A}}\right)\right\}_{i=1}^{N_{\mathrm{A}}}, s_i^{\mathrm{A}}=\left(s_{i1}^{\mathrm{A}},s_{i2}^{\mathrm{A}},\cdots,s_{im_{\mathrm{A}}}^{\mathrm{A}}\right)^{\mathrm{T}}, \ell_i^{\mathrm{A}}=\left(x_i^{\mathrm{A}},y_i^{\mathrm{A}}\right)^{\mathrm{T}}$，$N_{\mathrm{A}}$ 表示参考位置的个数，ℓ_i^{A} 表示第 i-th 个参考节点的物理位置，m_{A} 表示共享锚节点的个数，s_{ij}^{A} 表示在第 i-th 个参考节点处接收到的第 j-th 个 SAP 的信号强度；当室内环境动态变化后，随机采样若干未知位置的 RSS 指纹，作为不含位置信息的随机采样测试数据 $X_{\mathrm{rand}}^{\mathrm{B}}$，$X_{\mathrm{rand}}^{\mathrm{B}}=\left\{s_i^{\mathrm{B}}\right\}_{i=1}^{N_{\mathrm{B}}}, s_i^{\mathrm{B}}=\left(s_{i1}^{\mathrm{B}},s_{i2}^{\mathrm{B}},\cdots,s_{im_{\mathrm{B}}}^{\mathrm{B}}\right)^{\mathrm{T}}$，$N_{\mathrm{B}}$ 表示随机采样的采样点数，m_{B} 表示锚节点的个数，由于 A 阶段和 B 阶段处于同一室内环境中，故 $m_{\mathrm{B}}=m_{\mathrm{A}}$。然后进行潜在空间对齐[21]。

本文采用将插值法和拟合法融合在一起的广义延拓插值算法，该算法充分利用延拓域 Ω_e' 的额

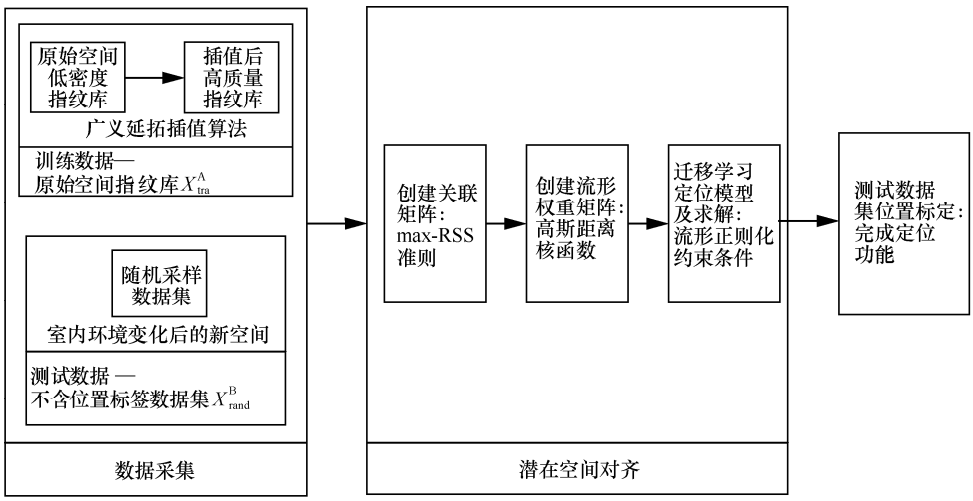


图 3　迁移学习动态定位算法示意图

外信息，使单元域内 Ω_e 的插值函数既能够充分利用邻近单元的信息，又保证插值函数与延拓域的逼近函数相互协调[23]，单元域和延拓域的划分如图 4 所示。

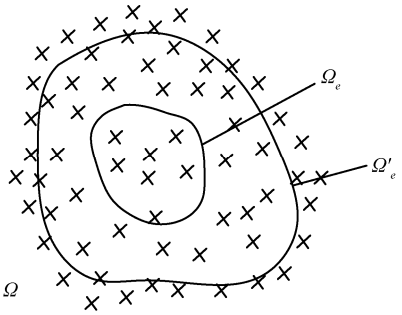


图 4　单元域及延拓域划分

广义延拓插值算法的数学模型如式（2）所示：

$$\begin{aligned}\min &= \sum_{j=1}^{v}\left[U^e\left(l_j\right)-u_{l_j}^q\right]^2 \\ &= \sum_{j=1}^{v}\left[\sum_{i=1}^{t}a_i^e\cdot h_i^e\left(l_j\right)-u_{l_j}^q\right]^2, l_j\in\Omega_e' \\ \text{s.t.}\ &\sum_{i=1}^{t}a_i^e\cdot h_i^e\left(l_j\right)=u_{l_j}^q, l_j\in\Omega_e, 1\leqslant q\leqslant M_{\text{A}}\end{aligned} \tag{2}$$

其中，$U^e\left(l_j\right)$ 表示拟合函数在位置 l_j 处的拟合值，v 是延拓域包含的数据个数，$u_{l_j}^q=\text{rss}_q^{l_j}$ 表示在位置 l_j 处测量的第 q 个锚节点的 RSS 值，t 是拟合函数的项数，$h_i^e\left(i=1,2,\cdots,t\right)$ 是延拓域 Ω_e' 上的一组基，$a_i^e\left(i=1,2,\cdots,t\right)$ 是待定系数[24]。

把各个单元子域拼接起来，作为整个可行域的拟合函数 $U(x)$：

$$U(x)=\sum_{e=1}^{m}U^e\left(x_e\right), x_e\in\Omega_e, x\in\Omega \tag{3}$$

本文采用二维正弦函数作为测试样本，验证广义延拓插值算法的精度，二维正弦函数为：

$$f(x,y)=\sin x+2\sin x\cos y+\cos y \tag{4}$$

其中，$x\in[0,10], y\in[0,5]$，原始测试函数以 0.5 步长进行采样，通过广义延拓插值算法和双线性插值算法反演出步长为 0.1 的采样曲面，并与真实曲面进行比较，表 1 统计了不同插值算法的插值误差概率为 80%和 90%时的插值精度。

表 1　不同算法插值精度

方法	CDF=80%	CDF=90%	平均误差
双线性插值法	0.06	0.08	0.056 8
广义延拓插值	0.03	0.04	0.031 6

从表 1 中可以看出，本文算法的插值精度明显优于双线性插值算法，本文算法可以通过稀疏

采样点还原出与真实函数变化趋势一致的高精度拟合曲面。故此，广义延拓插值算法可以在原始低工作量、低采样密度指纹库的基础上创建出高质量插值指纹库。

步骤 2　创建关联矩阵

为了实现 A 和 B 两个不同流形之间的嵌入对齐问题，需要创建类内关联图和类间关联图[25]。类内关联图表示矩阵 $X_{\text{tra}}^{\text{A}}$ 或 $X_{\text{rand}}^{\text{B}}$ 内部间的近邻特性，采用 k 近邻算法把每个 RSS 向量与其最近的邻居连接起来，如果矩阵中两个 RSS 向量在近邻范围内，则对应的关联矩阵元素值 g_{ij} 为 1，否则 g_{ij} 为 0，以此来构造类内关联矩阵 $G_{N_{\text{A}}\times N_{\text{A}}}^{\text{A}}$ 或 $G_{N_{\text{B}}\times N_{\text{B}}}^{\text{B}}$；类间关联图表示矩阵 $X_{\text{tra}}^{\text{A}}$ 和 $X_{\text{rand}}^{\text{B}}$ 数据间 RSS 相关对的近邻特性，矩阵 $X_{\text{tra}}^{\text{A}}$ 和 $X_{\text{rand}}^{\text{B}}$ 来自不同的特征空间，RSS 向量之间的距离不能直接计算，无法进行 RSS 相关对联接处理，考虑到同一室内环境中路径衰减因子在某个确定范围内取值，而环境噪声分布不同，需要先对原始数据进行噪声消除，然后利用 max-RSS 准则（对于每个共享锚节点 SAP，不同空间中该锚节点处接收到的最大信号强度通过 RSS 相关对的方式关联起来）[22]创建类间关联矩阵 $G_{N_{\text{A}}\times N_{\text{B}}}^{\text{AB}}$。

步骤 3　创建流形权重矩阵

根据关联矩阵和数据矩阵建立无向图的权重矩阵 W^{A}、W^{B}、W^{AB}，其中，权重矩阵 $W^{\text{A}}=\left[W^{\text{A}}(i,j)\right]_{N_{\text{A}}\times N_{\text{A}}}$ 采用高斯距离核函数创建[25]。

$$W^{\text{A}}(i,j)=\begin{cases}\exp\left(-\dfrac{\left\|s_i^{\text{A}}-s_j^{\text{A}}\right\|^2}{2\sigma_A^2}\right), & g_{ij}^{\text{A}}=1\\ 0, & \text{其他}\end{cases}\tag{5}$$

其中，$\|\ \|$ 表示欧氏距离，s_j^{A} 是数据矩阵中的列向量，$g_{ij}^{\text{A}}\in G^{\text{A}}$ 是关联矩阵中的元素，σ^{A} 是距离归一化因子，常选取数据矩阵的方差。

通过上述方法可以得到权重矩阵 W^{B}、W^{AB}，在实际应用中，由于 SAP 的个数远远小于采样点数，求得的 W^{AB} 往往是稀疏矩阵，无法把不同数据集间的 RSS 相关对有效地关联起来，因此利用迭代规则把相同数据集内部的 W^{A}、W^{B} 特性传播到 W^{AB} 中[22]，以此来增强不同数据集间 RSS 相关对的关联性。迭代规则如式（6）所示。

$$W_k^{\text{AB}}\leftarrow W^{\text{A}}\cdot W_{k\text{-}1}^{\text{AB}}\cdot W^{\text{B}}\tag{6}$$

其中，$k=1,2,\cdots,n$，是迭代的次数，W_0^{AB} 是通过高斯距离核函数创建的初始权重矩阵。通过迭代足够的次数，使 W^{AB} 矩阵能有效地把两个数据集的潜在特性关联起来。

步骤 4　迁移学习定位模型及求解

设 $P^K=\left\{p_i^{(K)}\right\}_{i=1}^{N_{\text{A}}+N_{\text{B}}}\in\mathbb{R}^{K\times(N_{\text{A}}+N_{\text{B}})}$，$p_i^{(K)}=\left(p_i^1,p_i^2,\cdots,p_i^K\right)^{\text{T}}$ 表示训练数据和测试数据在 K 维潜在空间的嵌入坐标矩阵。当 $i=1,2,\cdots,N_{\text{A}}$ 时，$p_i^{(K)}$ 表示训练数据集 $X_{\text{tra}}^{\text{A}}$ 中第 i 个 RSS 向量 s_i^{A} 在 K 维潜在空间中的嵌入坐标，当 $j=N_{\text{A}}+1,N_{\text{A}}+2,\cdots,N_{\text{A}}+N_{\text{B}}$ 时，$p_j^{(K)}$ 表示测试数据集 $X_{\text{rand}}^{\text{B}}$ 中第 j 个 RSS 向量 s_j^{B} 在 K 维潜在空间中的嵌入坐标。为了让嵌入坐标矩阵在整个空间中保持平滑，需要增加流形正则化的约束条件[26]，其定位模型如式（7）所示。

$$\begin{aligned}\widetilde{P^K}&=\min_{P^K}\left(\sum_{i=1}^{N_{\text{A}}}\sum_{j=1}^{N_{\text{B}}}\left(P_i^{(K)}-P_{N_{\text{A}}+j}^{(K)}\right)^2\cdot W^{\text{AB}}(i,j)\right)\\&=\min_{P^K}\ P^K\cdot L^{\text{AB}}\cdot\left(P^K\right)^{\text{T}}\\&\text{s.t.}\ \left(P_i^{(K)}\right)^{\text{T}}\cdot P_i^{(K)}=1,\ i=1,2,\cdots,N_{\text{A}}+N_{\text{B}}\end{aligned}\tag{7}$$

其中，L^{AB} 是权重矩阵 W^{AB} 的拉普拉斯矩阵[27]，如式（8）所示。

$$\begin{cases}L^{\text{AB}}=(1-\mu)\cdot\text{diag}\left(L^{\text{A}},L^{\text{B}}\right)+\mu\cdot L_{(N_{\text{A}}+N_{\text{B}})\times(N_{\text{A}}+N_{\text{B}})}^{\text{AB}}\\ L_{(N_{\text{A}}+N_{\text{B}})\times(N_{\text{A}}+N_{\text{B}})}^{\text{AB}}=\begin{bmatrix}D_{N_{\text{A}}\times N_{\text{A}}}^{\text{AB}} & -W^{\text{AB}}\\ -\left(W^{\text{AB}}\right)^{\text{T}} & D_{N_{\text{B}}\times N_{\text{B}}}^{\text{BA}}\end{bmatrix}\end{cases}\tag{8}$$

其中，L^{A}、L^{B} 分别是矩阵 W^{A}、W^{B} 的拉普拉斯矩阵，$L^{\text{A}}=D^{\text{A}}-W^{\text{A}}$，矩阵 D^{A} 是对角矩阵，对角线上的元素 $D^{\text{A}}(i,i)=\sum_{j=1}^{N_{\text{A}}}W^{\text{A}}(j,i)$，$D_{N_{\text{A}}\times N_{\text{A}}}^{\text{AB}}$ 是对角矩阵，对角线上的元素为 $D^{\text{AB}}(i,i)=\sum_{j=1}^{N_{\text{B}}}W^{\text{AB}}(i,j)$，

$D^{\mathrm{BA}}_{N_\mathrm{B}\times N_\mathrm{B}}$ 是对角矩阵，对角线上的元素为 $D^{\mathrm{BA}}(i,i)=\sum_{j=1}^{N_\mathrm{A}} W^{\mathrm{AB}}(j,i)$，$0\leqslant\mu\leqslant1$ 是类内数据与类间数据间的归一化因子，取值越大代表类间数据的权重越高。

迁移学习定位模型式（7）是一个典型的数学优化问题，根据最优化理论可知该定位模型是求解最小特征值问题[27]，即求解出矩阵 L^{AB} 的前 K 个最小特征值（不包括 0 特征值）对应的特征向量，作为 P^K 的最优估计 $\widetilde{P^K}$。

步骤 5 完成定位

如上所述，$\widetilde{P^K}$ 的前 N_A 个列向量是训练数据 $X^{\mathrm{A}}_{\mathrm{tra}}$ 在 K 维潜在空间中的嵌入坐标，$\widetilde{P^K}$ 的后 N_B 个列向量是测试数据集 $X^{\mathrm{B}}_{\mathrm{rand}}$ 的嵌入坐标，然后根据 k 近邻或加权 k 近邻算法实现未标签数据集 $X^{\mathrm{B}}_{\mathrm{rand}}$ 位置的自动标定。

潜在空间维度的确定通常有两种方法，即含有位置约束和不含位置约束[22]。含有位置约束的定位模型是把测试数据集的位置信息作为约束条件，并将不同类之间的数据对齐到二维潜在空间中；不含位置约束是测试数据集的位置信息仅仅参与位置标签标定过程而不参与嵌入对齐过程，此时空间维数采用式（9）的方法进行确定[28]。

$$\frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{n}\lambda_i}\leqslant\xi^* \tag{9}$$

其中，ξ^* 是潜在空间保留信息的阈值，d 是潜在空间的维数，λ_i 是从小到大排列的第 i 个非 0 特征值。

4 迁移学习动态定位性能仿真与测试

4.1 广义延拓插值指纹库测试及性能仿真

为得到实际环境中的 RSS 距离衰减模型，本文在如图 5（a）的车库环境中进行实际测试，采用 3 对 CC2530 芯片的 ZigBee 开发板在同一场景中每隔 0.2 m 进行 RSS 数据采集。

(a) 室内车库原始环境　　(b) 车库环境变化以后

图 5　测试车库环境

本文算法实验数据均来自于真实的环境，通过数据拟合得到 3 对锚节点的 RSS 距离衰减曲线，如式（10）所示，在同一个静态室内环境中，3 条衰减曲线相差并不是很大，因此可认为该车库环境中，RSS 距离衰减模型为式（11）：

$$\begin{cases}\mathrm{RSS}_{1,\mathrm{dB}}=-12.17\times\lg d-48.63+X_{1,\sigma}\\ \mathrm{RSS}_{2,\mathrm{dB}}=-11.53\times\lg d-45.72+X_{2,\sigma}\\ \mathrm{RSS}_{3,\mathrm{dB}}=-11.61\times\lg d-46.12+X_{3,\sigma}\end{cases} \tag{10}$$

$$\mathrm{RSS}_{\mathrm{dB}}=-11.73\times\lg d-46.68+X_\sigma \tag{11}$$

其中，$X_\sigma\sim\mathcal{N}(0.001,15.468)$ 是高斯白噪声。

本文在 MATLAB 2014b 软件中对广义延拓插值算法的性能进行仿真，其中，RSS 传播模型采用式（11）实测的参数，在车库环境为 20 m×15 m 范围内，13 个锚节点分布如图 6(a)所示。原始 RSS 指纹库是通过每隔 2 m 采集 RSS 指纹建立的，此时原始指纹库的单位面积内的采样密度为 0.25，广义延拓插值算法的插值距离是 0.5 m，故广义延拓插值 RSS 指纹库的采样密度为 4，相比较原始指纹库采样密度提高了 16 倍。其中含有室内环境噪声与不含环境噪声的插值算法的仿真结果如图 6(b)所示，其中室内环境噪声是服从 $X_\sigma\sim\mathcal{N}(0.001,15.468)$ 分布的高斯白噪声。通过图 6(b)可知，通过与真实 RSS 指纹库相比较，插值以后的 RSS 误差在 2 dBm、5 dBm 和 10 dBm 精度以内的置信概率分别为 70%、82%和 95%，说明该算法通过对低采样密度的原始指纹库进行插值，可形成与真实指纹库接近的高质量的 RSS 指纹库，故此广义延拓插值算法可以节省大量的人力和财力资源。

4.2 模型参数对定位性能的影响

本文提出的迁移学习动态定位模型可以有效

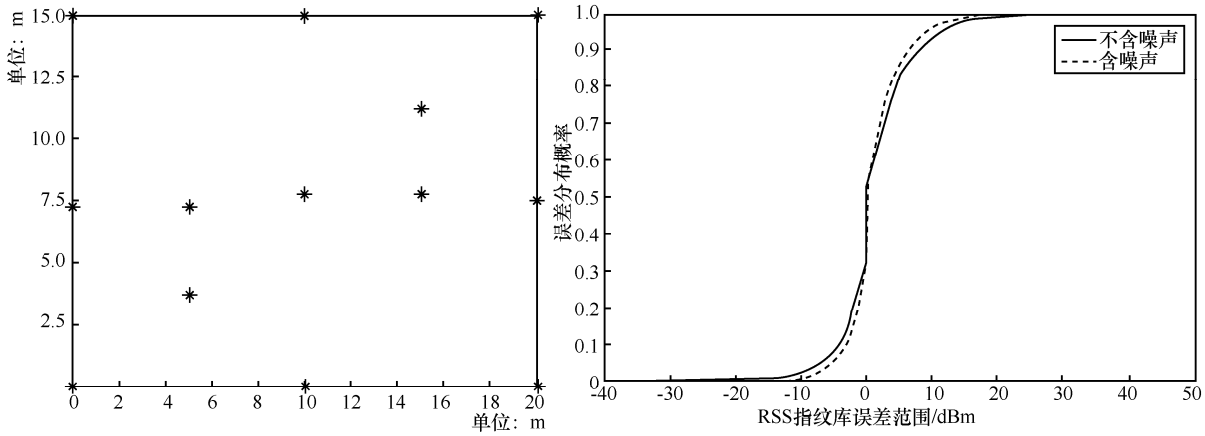


图 6　室内锚节点分布及广义延拓插值算法性能仿真

解决室内环境发生动态变化以后的定位问题，该模型由很多关键参数构成，例如室内锚节点 AP 的个数、潜在空间的保留信息阈值 ξ^*、完成定位时 k 近邻或加权 k 近邻算法的近邻个数、高斯距离核函数的距离归一化因子 σ^{A} 以及迭代规则中的迭代次数等。故此本文主要对上述前 3 个关键参数进行研究，分别讨论这些参数对迁移学习动态定位性能的影响。

4.2.1　室内锚节点 AP 个数对定位性能的影响

图 7 是平均定位误差 RMSE 随室内锚节点 AP 个数的变化情况曲线，随着 AP 数量的增加，迁移学习动态定位算法与 k 近邻、加权 k 近邻定位算法的平均定位误差逐渐变小。但是定位误差并不是随着 AP 数量增加单调减小的，因为待定位节点可能会选取距离较远的 AP，此时锚节点距离较远，信号较弱，RSS 数值较小，受环境影响严重，不利于定位。

4.2.2　潜在空间的保留信息阈值对定位性能的影响

潜在空间的保留信息阈值 ξ^* 是一个在(0,1)范围内取值的常数，由式（9）可知，阈值代表潜在对齐空间的维度，当阈值 ξ^* 确定以后，室内环境发生动态变化后计算出来的潜在空间的维数是不固定的，这种动态调整维数的方法避免了使用固定维数带来的问题。图 8 是潜在空间的

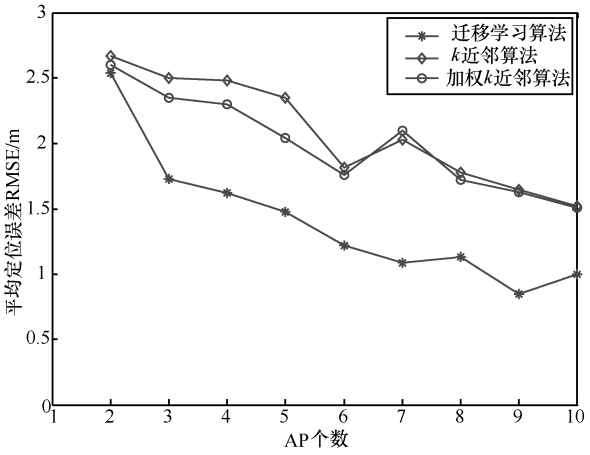


图 7　锚节点 AP 个数对定位性能的影响

保留信息阈值对定位性能的影响，从图 8 中可知，当阈值 ξ^* 比较小时，相当于潜在空间的维数较小，无法准确地把环境变化后的潜在信息进行有限映射，此时定位误差比较大，想要达到高精度定位比较困难；随着阈值 ξ^* 的增大，位置结算所需的定位潜在空间维数增加，平均定位误差会减小，但是当阈值 ξ^*=0.3 左右时，平均定位误差最小；阈值 $\xi^* > 0.8$ 后，潜在空间维数较大，会引入更多无用的位置指纹信息，造成平均定位误差变大。因此可以把 ξ^*=0.3 作为最佳的潜在空间维数的阈值。

4.2.3　定位时 k 近邻的近邻个数对定位性能的影响

参数 k 是迁移学习动态定位算法完成定位时 k

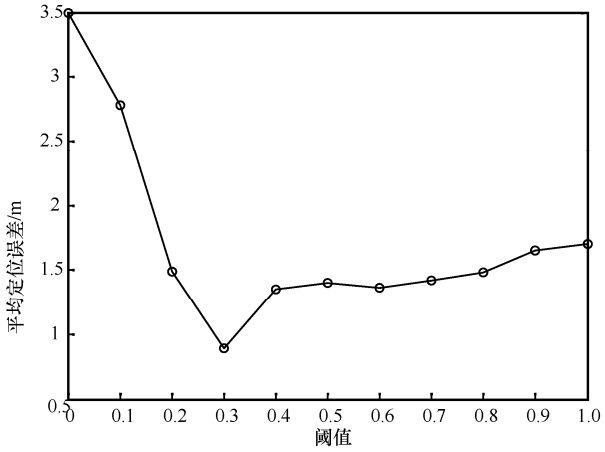


图 8　潜在空间的保留信息阈值对定位性能的影响

近邻算法所使用的近邻数，当 AP=8 时，平均定位误差与 k 之间的关系如图 9 所示。随着近邻个数 k 值变大，参与定位的参考节点近邻个数变多，系统的平均定位误差变小。但经过大量仿真发现，k 值并非越大越好，如果 k 值很大，会把距离待定位目标节点较远的指纹点纳入近邻数中，在求解位置坐标时这些距离较远的指纹往往是无效的信息，故此会影响系统整体的定位精度。从图 9 中可以看出 k 取 6 时平均定位误差最小，可作为最佳的 kNN 近邻个数。

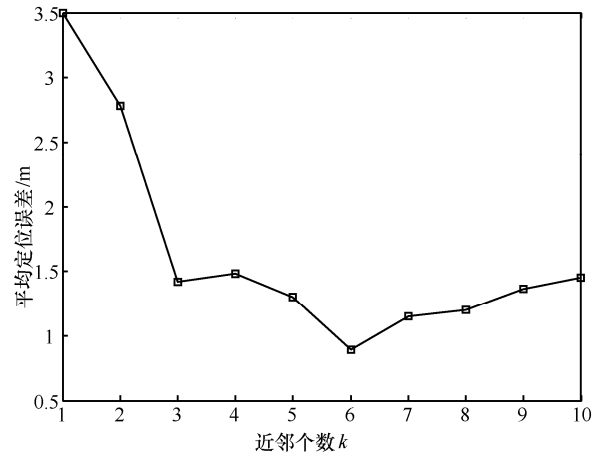


图 9　近邻个数 k 对定位性能的影响

4.3　迁移学习定位性能仿真

为了验证本文所提出的迁移学习动态定位算法的定位性能，本文所选的室内环境与图 5(a)中是同一个车库，不同的是车库中停放的车辆发生明显改变，如图 5(b)所示，经过测试得到该环境下 RSS 衰减模型如式（12）所示，其中室内车库变化前后地 RSS 衰减曲线如图 10 所示。从图 10 可知，当环境变化后，RSS 衰减得更厉害。室内锚节点 AP 的个数为 13 个，潜在空间的保留信息阈值 ξ^*=0.3，完成定位时 k 近邻或加权 k 近邻算法的近邻个数为 6 个，距离归一化因子 σ^{A} 选取训练数据 $X_{\mathrm{tra}}^{\mathrm{A}}$ 或者测试数据 $X_{\mathrm{rand}}^{\mathrm{B}}$ 的方差，迭代规则中的迭代次数为 3 次。本文采用传统 kNN 算法[15]处理环境变化后的指纹定位系统作为对比方案，定位性能仿真结果如图 11 所示，其中不同算法的定位精度概率分布见表 2。

$$\mathrm{RSS}_{\mathrm{dB}} = -16.25 \times \lg d - 60.35 + \mathcal{N}(0,15.283) \quad (12)$$

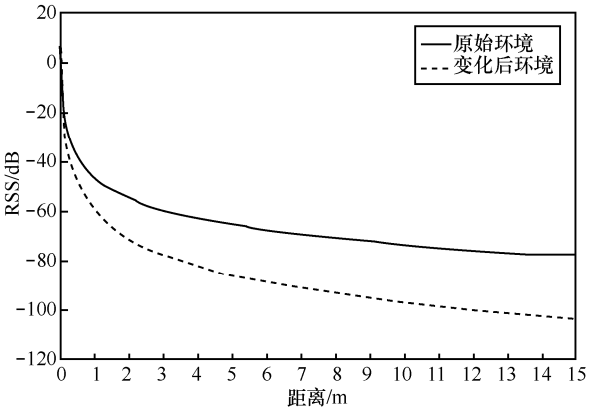


图 10　车库环境变化前后 RSS 衰减曲线

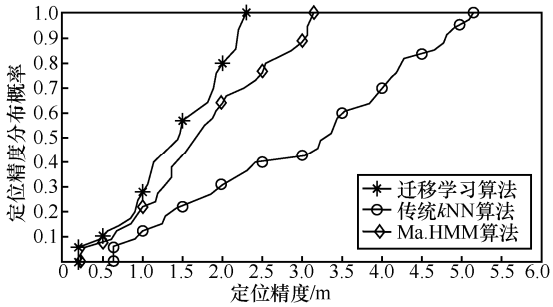


图 11　不同算法定位误差累积分布曲线

从图 11 和表 2 可以看出，迁移学习定位算法在 0.5 m、1 m 和 1.5 m 内的定位精度置信概率分别为 9.5%、29.1%和 57.6%，分别高出传统 kNN 算法 9.5%、16.8%和 33.8%，高出 Ma.HMM 算法 0.5%、8.8%和 15.2%，并且迁移学习算法的平均定位精度是 1.23 m，传统 kNN 算法[15]的平均定位精

表 2　不同算法定位精度的概率分布值

方法	0.5 m	1.0 m	1.5 m	2.0 m	2.5 m	平均精度/m
*k*NN 算法[15]	0	12.3%	23.8%	30.3%	40.7%	3.28
Ma.HMM 算法[12]	9.0%	20.3%	42.4%	64.5%	77.5%	1.74
迁移学习算法	9.5%	29.1%	57.6%	80.5%	100%	1.23

度为 3.28 m，Ma.HMM 算法[12]的平均定位精度为 1.74 m，迁移学习定位算法的定位误差有 57.6%的概率在 1.5 m 以内，而传统机器学习算法的定位误差高达 5 m。这说明当室内环境特征发生较大改变的时候，本文所提出的迁移学习动态定位算法和基于隐性马尔可夫模型的 Ma.HMM 定位算法[12]的性能更加优越，且比传统 *k*NN 算法更有利于实现小误差高精度定位（平均定位误差在 1.2 m 左右）。故此本算法无论在定位精度、定位误差范围还是小误差定位的置信概率方面，相比传统机器学习定位算法都具有明显的优势。

通过上述实验可知，室内环境特征发生明显改变时，传统机器学习指纹定位算法已经无法进行室内定位，Ma.HMM 算法虽然能够进行迁移定位但是定位误差较大，而本文所提出的迁移学习动态定位算法具有较小的定位误差范围和较高的定位精准度。

5　结束语

为了消除室内环境动态变化对系统定位性能造成的不利影响，本文设计了一种迁移学习动态定位算法，通过在环境变化以后的空间中随机采样，找到潜在的特征空间进行位置的标定功能，从而避免了原始指纹库失效造成的定位精度降低的问题。此外，本文提出的定位算法还需要进一步改善，特别是动态定位问题，可以采用流形对齐、深度学习等人工智能算法，充分利用定位节点之间的空间关联性，进一步提升定位系统的定位精度，同时系统应该具有判断异常数据的能力，以保障系统的定位质量和人性化体验。基于此，上述定位技术目前也处于研究阶段。

参考文献：

[1] LUO C, CHENG L, CHAN M C, et al. Pallas: self- bootstrapping fine-grained passive indoor localization using WiFi monitors[J]. IEEE Transactions on Mobile Computing, 2017, PP(99): 1.

[2] 朱亚萍, 夏玮玮, 章跃跃, 等. 基于 RSSI 和惯性导航的融合室内定位算法[J]. 电信科学, 2017, 33(10): 99-106.
ZHU Y P, XIA W W, ZHANG Y Y, et al. A hybrid indoor localization algorithm based on RSSI and inertial navigation[J]. Telecommunications Science, 2017, 33(10): 99-106.

[3] 徐小良, 高健, 黄河, 等. 基于 RSS 空间线性相关的 WLAN 位置指纹定位算法[J]. 电信科学, 2017, 33(3): 14-21.
XU X, GAO J, HUANG H, et al. Fingerprint localization algorithm based on linear spatial dependence of WLAN RSS[J]. Telecommunications Science, 2017, 33(3): 14-21.

[4] 尚俊娜, 程涛, 盛林, 等. 广义延拓插值模型在 RSSI 测距方法中的应用[J]. 传感技术学报, 2016, 29(11).
SHANG J N, CHENG T, SHENG L, et al. Application of generalized extended interpolation method in distance measurement based on RSSI[J]. Chinese Journal of Sensors and Actuators, 2016, 29(11).

[5] HASSAN-ALI M, PAHLAVAN K. A new statistical model for site-specific indoor radio propagation prediction based on geometric optics and geometric probability[J]. IEEE Transactions on Wireless Communications, 2002, 1(1): 112-124.

[6] KING T, KOPF S, HAENSELMANN T, et al. COMPASS: a probabilistic indoor localization system based on 802.11 and digital compasses[C]//ACM Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization, September 29, 2006, Los Angeles, CA, USA. New York: ACM Press, 2006: 34-40.

[7] YOUSSEF M A, AGRAWALA A, SHANKAR A U. WLAN location determination via clustering and probability distributions[C]//IEEE International Conference on Pervasive Computing and Communications, March 23-26, 2003, Fort Worth, TX, USA. Piscataway: IEEE Press, 2003: 143-150.

[8] 张伟, 花向红, 邱卫宁, 等. Wi-Fi 指纹定位的一种新组合算法[J]. 测绘工程, 2017, 26(3): 14-18.
ZHANG W, HUA X H, QIU W N, et al. A new combinatorial optimization algorithm for WiFi positioning[J]. Engineering of Surveying and Mapping, 2017, 26(3): 14-18.

[9] 张宏刚, 黄华. 基于 RSSI 路径损耗因子动态修正的三边质心定位算法[J]. 传感技术学报, 2016, 29(11).
ZHANG H G, HUANG H. Dynamic correction algorithm for

trilateral centroid localization based on RSSI path loss factor[J]. Chinese Journal of Sensors and Actuators, 2016, 29(11).

[10] 张梦丹, 卢光跃, 王宏刚, 等. 基于线性内插法改进的室内定位算法[J]. 电信科学, 2017, 33(1): 9-15.
ZHANG M D, LU G Y, WANG H G, et al. An improved indoor location algorithm based on linear interpolation[J]. Telecommunications Science, 2017, 33(1): 9-15.

[11] SOROUR S, LOSTANLEN Y, VALAEE S, et al. Joint indoor localization and radio map construction with limited deployment load[J]. IEEE Transactions on Mobile Computing, 2015, 14(5): 1031-1043.

[12] CAI X, CHEN L, CHEN G. Constructing adaptive indoor radio maps for dynamic wireless environments[R]. 2013.

[13] 陈斌涛, 刘任任, 陈益强, 等. 动态环境中的 WiFi 指纹自适应室内定位方法[J]. 传感技术学报, 2015(5): 729-738.
CHEN B T, LIU R R, CHEN Y Q, et al. WiFi fingerprint based self-adaptive indoor localization in the dynamic environment[J]. Chinese Journal of Sensors and Actuators, 2015(5): 729-738.

[14] YIN J, YANG Q, NI L M. Learning adaptive temporal radio maps for signal-strength-based location estimation[J]. IEEE Transactions on Mobile Computing, 2008, 7(7): 869-883.

[15] 孔港港, 杨力, 孙聃石, 等. 一种基于位置指纹定位的 *K*-均值聚类算法的改进[J]. 全球定位系统, 2016, 41(5): 89-92.
KONG G G, YANG L, SUN D S, et al. Research on an algorithm of fingerprint location based on *K*-means and W*K*NN[J]. GNSS World of China, 2016, 41(5): 89-92.

[16] 戴文渊. 基于实例和特征的迁移学习算法研究[D]. 上海: 上海交通大学, 2008.
DAI W Y. Instance-based and feature-based transfer learning[D]. Shanghai: Shanghai Jiao Tong University, 2008.

[17] ALI S, NOBLES P. A novel indoor location sensing mechanism for IEEE 802.11 b/g wireless LAN[C]//The Workshop on Localization, March 22, 2007, Hannover, Germany. Piscataway: IEEE Press, 2007: 9-15.

[18] SUN Z, CHEN Y, QI J, et al. Adaptive localization through transfer learning in indoor Wi-Fi environment[C]//Seventh International Conference on Machine Learning and Applications. IEEE Computer Society, Dec 13-18, 2008, San Diego, CA, USA. Piscataway: IEEE Press, 2008: 331-336.

[19] ZHENG V W, XIANG E W, YANG Q, et al. Transferring localization models over time[C]// AAAI Conference on Artificial Intelligence, AAAI 2008, July 13-17, 2008, Chicago, Illinois, USA. New York: ACM Press, 2008: 1421-1426.

[20] DAI W, CHEN Y, XUE G R, et al. Translated learning: transfer learning across different feature spaces[C]//Conference on Neural Information Processing Systems, December 8-10, 2008, Vancouver, British Columbia, Canada. New York: ACM Press, 2008: 353-360.

[21] PAN S J, SHEN D, YANG Q, et al. Transferring localization models across space[C]// National Conference on Artificial Intelligence, July 13-17, 2008, Chicago, Illinois. New York: ACM Press, 2008: 1383-1388.

[22] WANG H Y, ZHENG V W, ZHAO J, et al. Indoor localization in multi-floor environments with reduced effort[C]//IEEE International Conference on Pervasive Computing and Communications, March 29-April 2, 2010, Mannheim, Germany. Piscataway: IEEE Press, 2010: 244-252.

[23] 施沛立, 颜毅华, 徐国华. 工程科学中的广义延拓逼近法(精)[M]. 北京: 科学出版社, 2005.
SHI H L, YAN Y H, XU G H. Generalized continuation approximation method in engineering science(fine)[M]. Beijing: Science Press, 2005.

[24] 张志富, 裴军, 胡超, 等. 车载动中通信标跟踪的广义延拓逼近算法[J]. 西安电子科技大学学报(自然科学版), 2017, 44(4): 112-117.
ZHANG Z F, PEI J, HU C, et al. Generalized extended approximation algorithm for vehicle satellite communication on the move[J]. Journal of Xidian University(Natural Science), 2017, 44(4): 112-117.

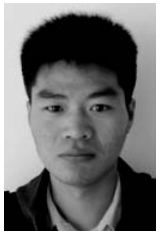
[25] ZHOU C F, MA L, TAN X Z. Joint semi-supervised RSS dimensionality reduction and fingerprint based algorithm for indoor localization[R]. 2014.

[26] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples[J]. Journal of Machine Learning Research, 2006, 7(1): 2399-2434.

[27] BELKIN M, NIYOGI P. Laplacian Eigenmaps for dimensionality reduction and data representation[M]. Cambridge: MIT Press, 2003.

[28] FANG S H, LIN T. Principal component localization in indoor WLAN environments[J]. IEEE Transactions on Mobile Computing, 2011, 11(1): 100-110.

[作者简介]



刘参（1990−），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为无线传感器网络、无线定位等。

尚俊娜（1979−），女，博士，杭州电子科技大学通信工程学院副教授，主要研究方向为通信信号处理、智能算法。

李蕊江（1993−），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为信号处理、无线定位等。

岳克强（1984−），男，博士，杭州电子科技大学电子信息学院讲师，主要研究方向为进化计算、通信信号处理。

研究与开发

基于业务类型的集中式接入网基站处理资源分配算法

张新苹[1,2]，王园园[2]，田霖[2]，郝树良[1]
（1. 重庆邮电大学，重庆 400065；2. 中国科学院计算技术研究所，北京 100190）

摘　要：现有的资源分配算法研究主要面向数据中心和云计算环境展开，未能充分考虑基站处理资源的多样性（CPU、内存、网络带宽、FPGA、DSP）和业务种类的多样性，不能直接应用于集中式接入网架构中。针对该问题，提出了基于业务类型的资源分配算法，首先采用 Fisher 分割方法，根据用户业务类型对不同类型计算资源的需求，对业务进行分类，然后利用资源分配均衡策略分配基站处理资源。仿真结果表明，该算法有效地减少了开启物理服务器的个数并提高了物理服务器的资源利用率，达到了绿色节能的目的。

关键词：集中式接入网；基站处理资源分配；用户业务类型；Fisher 分割方法；资源分配均衡策略
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018202

Service aware base station processing resource allocation for centralized radio access network

ZHANG Xinping[1,2], WANG Yuanyuan[2], TIAN Lin[2], HAO Shuliang[1]
1. Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Current research on resource allocation algorithms mainly focuses on data centers, and cloud computing environments. These resource allocation algorithms did not consider the diversity of processing resources (CPU, memory, network bandwidth, FPGA, DSP) and the diversity of service types in the centralized base station, resulting in the low processing resource utilization. In order to solve this problem, a service oriented base station processing resource allocation algorithm was proposed. Firstly, the Fisher partition method was used to classify the base station's processing resource requirements according to the user's service request. Then, the resource allocation balancing strategy was used to allocate the base station processing resources. Experimental results show that the algorithm is effective to reduce the number of physical servers and improves the resource utilization of the servers, realizing energy conservation and communications.

Key words: centralized radio access network, base station processing resource allocation, user service type, Fisher partition method, resource allocation balancing strategy

收稿日期：2017−12−19；**修回日期**：2018−06−08
基金项目：国家自然科学基金资助项目（No.61431001）
Foundation Item：The National Natural Science Foundation of China (No.61431001)

1 引言

近年来，蜂窝移动通信系统发展迅猛，网络容量需求呈指数级增长[1,2]。传统蜂窝移动通信系统采用大规模部署基站的方式提升地区覆盖容量，各基站之间相对独立，基站处理资源按照基站覆盖区域内用户业务峰值容量部署，且资源独占。这种组网架构使得孤立的基站不能处理具有“潮汐效应”的动态网络负载，当负载降低时，基站处理资源的利用率就会明显降低，资源浪费严重。为了解决传统架构的这些问题，业界提出了集中式接入网的概念和架构，例如中国移动的C-RAN（centralized, cooperative, cloud RAN）[3]和中国科学院计算技术研究所（以下简称中科院计算所）的物理集中、逻辑分布的超级基站架构[4]。在集中式接入网中，基站处理资源集中部署形成资源池，基于资源水平共享、统计复用，通过资源管控实现处理资源的动态分配[5-7]。然而集中统一地管理所有处理资源必然给系统管理和资源分配带来新的挑战。如何根据业务实际负载需求，对基站处理资源进行灵活的分配，提高基站处理资源的利用率，是集中式接入网架构的一个重要研究点。

目前，大规模处理资源分配算法的研究主要集中在集中式数据中心、云计算中心等环境下。云计算资源调度通常以虚拟机为资源分配单位[8-10]，将虚拟机调度到一个或多个物理机上，研究的是虚拟机和物理机之间的映射关系。大量的研究将这种问题归结为常见的装箱问题[11-12]，求解目标是用尽可能少的物理资源来满足所有虚拟机的资源需求。参考文献[13-14]在CPU、内存资源约束情况下建立模型，以提高资源利用率为目标，通过遗传算法求解多约束优化问题，获得物理机和虚拟机的优化映射方案。但是采用启发式算法求的只是局部最优解，时间复杂度较高，可扩展性和算法实时性不强。参考文献[15-16]提出了基于业务量预测的资源分配算法。云服务提供商首先对云用户的业务量进行预测，然后按照所预测的业务量为相应的云用户预先配置虚拟机资源，存在的不足是当突发业务到来时会出现预测业务量不准确的情况，这就会导致预先配置的虚拟机资源过大或过小，从而造成资源利用率较低或不满足业务需求的问题。由于无线接入网所需的资源类型、处理的用户业务类型等与云计算场景都有较大不同，因此云计算中的资源调度算法虽然具有一定的借鉴意义，但不能直接用于接入网中。在集中式接入网架构下，基站处理资源分配算法主要考虑以下两点：

- 集中式接入网的基站具备类型多样的计算资源，包括CPU、内存、网络带宽、FPGA、DSP等，其中CPU、内存、网络带宽主要完成数据处理和传输；DSP、FPGA作为基带处理的加速器，主要作用是快速地实现各种数字信号处理算法，满足通信系统的实时性要求。
- 通信系统中用户业务的多样性决定了对资源的需求量不同[17]，如对于计算类业务（用户数据处理），它们对计算资源的需求较高，而对于存储类任务（下载音频或视频），它们对存储和带宽资源的需求较高。

针对以上两点，本文提出了基于业务类型的基站处理资源分配算法（station processing resource allocation algorithm，SPRAA），主要思想是在集中式接入网架构下进行基站处理资源分配时，根据用户业务类型对不同类型计算资源的需求，将处理资源需求互补的不同业务组合到一起，按业务组进行处理资源的分配，有助于提高资源利用率。本文以减少物理服务器个数和提高资源利用率为目标对上述问题进行数学建模，通过Fisher业务分类和资源分配均衡策略来完成处理资源的灵活分配，达到提高资源利用率的目的。

2　处理资源分配算法的设计与实现

2.1　算法设计与实现

2.1.1　业务类型分类

在某一个时间窗内到达的业务种类是多样的，不同类型的业务，其数据处理对不同类型处理资源的需求量也会不同[18-19]。为了降低资源分配的复杂度，首先对到达的业务进行分类，分类标准是业务对基站处理资源（CPU、内存、带宽资源等）的需求大小。为了把业务分成 K 类，同时使得业务分类结果更加精确，避免出现“步骤 4 根据资源占重比求得业务类型”中业务类型不确定的情况，比如当各类资源占重比为（0.4，0.1，0.4，0.1）时，CPU 和带宽的占重比是相同的，确定业务类型时会出现分歧。考虑到业务分类结果的精确性，分类方法采取参考文献[20]的 Fisher 最优分割法，分类步骤如下。

步骤 1　数据标准化

设基站处理资源类型为 P 项，包括 CPU、内存、带宽、DSP、FPGA 等资源类型。在某个时间窗内到达的业务个数为 N 个，业务集合为 $T=\left[T_1,T_2,T_3,\cdots,T_N\right]$，对处理资源的需求矩阵为：

$$X_{(P\times N)}=\begin{bmatrix} x_{11} & x_{12}\cdots & x_{1N} \\ x_{21} & x_{22}\cdots & x_{2N} \\ \vdots & \vdots & \ddots \vdots \\ x_{P1} & x_{P2}\cdots & x_{PN} \end{bmatrix} \tag{1}$$

其中，元素 x_{ij} 表示第 j 个业务对第 i 类计算资源的需求值。为了减小数据的计算量，采用 min-max 标准化（min-max normalization）对矩阵 X 进行处理得到矩阵 Z，使得 Z 中每个元素 z_{ij} 大小在 $[0,1]$ 区间内。定义矩阵 Z 中每个 z_{ij} 的计算式如下：

$$z_{ij}=\frac{x_{ij}-\min_{1\leqslant j\leqslant N}\left\{x_{ij}\right\}}{\max_{1\leqslant j\leqslant N}\left\{x_{ij}\right\}-\min_{1\leqslant j\leqslant N}\left\{x_{ij}\right\}} \tag{2}$$
$$\left(i=1,2,\cdots P;j=1,2,\cdots,N\right)$$

得到矩阵 $Z_{(P\times N)}=[z_{ij}]$。

步骤 2　计算极差矩阵

在不打乱业务顺序的情况下，把某个周期内到达的 N 个业务进行分割（分类），其所有可能的分割共有 $E=C_{N-1}^1+C_{N-1}^2+\cdots\cdots+C_{N-1}^{N-1}=2^{N-1}$ 种划分方法。在所有的分割方法中，存在一种分割使得同一类内的所有业务对各类计算资源的需求参数之间的差异最小，而使不同类业务间的资源需求差异最大。在同一类内数据的相似性用极差 d_{ij} 表示，其定义如下：

$$d_{ij}=\begin{cases}\sum_{a=1}^{P}(\max_{i\leqslant b\leqslant j}\boldsymbol{Z}_{ab}-\min_{i\leqslant b\leqslant j}\boldsymbol{Z}_{ab}),\ i\leqslant j\\ 0,i>j\end{cases} \tag{3}$$

其中，d_{ij} 表示业务 $T=\left[T_i,T_{i+1},T_{i+2},\cdots,T_j\right]$ 对基站处理资源参数的相似性大小，d_{ij} 越小，表示这类业务对所需要的处理资源越相似。由式（3）得到矩阵 D：

$$D=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ & d_{22} & \cdots & d_{2N} \\ & & & \vdots \\ & & \cdots & d_{NN} \end{bmatrix} \tag{4}$$

步骤 3　最优 K 分割

用 $b(n,k)(n\geqslant k)$ 表示 n 个业务分为 k 类的某一种分法，判断某种分法的好坏用分类损失函数 $L\left[b(n,k)\right](n\geqslant k)$ 来衡量。当 n 和 k 固定时，$L\left[b(n,k)\right]$ 越小表示分类越合理。因此要寻找一种分法 $b(n,k)$，使分类损失函数 $L\left[b(n,k)\right]$ 达到最小，递推计算式如下：

$$\begin{cases}L\left[b(n,2)\right]=\min_{2\leqslant j\leqslant n}\left\{d_{1j-1}+d_{jn}\right\}\\ L\left[b(n,k)\right]=\min_{2\leqslant j\leqslant n}\left\{L\left[b(j-1,k-1)\right]+d_{jn}\right\}\end{cases} \tag{5}$$

当分类数 k 确定时，记分割点为 $(j_2,j_3,j_4,\cdots,j_K)$。要想找到分割点 $j_k(2\leqslant k\leqslant K)$，即要找到将业务基于处理资源需求参数分为 2 类、3 类 $\cdots$ K 类的分割点，依据式（5）求出 $L\left[b(n,k)\right](2\leqslant k\leqslant K)$ 的最小值，最小值对应的 j 即分割点。例如当 k=2 时，求得 $L\left[b(n,2)\right]$ 最小

值，其对应的 j 即分割点 j_2；k=3 时，求得 $L\left[b(n,3)\right]$ 最小值，其对应的 j 即分割点 j_3；以此类推直到求得分割点 j_K。

基于以上分析根据业务的处理资源需求参数对业务进行最优二分割：根据矩阵 D 计算将用户业务分两类的各种分割相应的损失函数 $L\left[b(j,2)\right](j=2,3,4\cdots,n)$：

$$\begin{aligned} &L\left[b(2,2)\right]=d_{11}+d_{22}\\ &L\left[b(3,2)\right]=\min(d_{11}+d_{23},\ d_{12}+d_{33})\\ &\quad\vdots\\ &L\left[b(n,2)\right]=\min_{2\leqslant j\leqslant n}(d_{1j-1}+d_{jn}) \end{aligned} \tag{6}$$

找到 $L\left[b(j,2)\right](j=2,3,4,\cdots,n)$ 的最小值，则分割点的位置是 j。由式（5）可以求得最优三分割、…一直到 K 分割。在某个周期内，通过 Fisher 最优分割法，将到达的业务分成 K 类，第 $k(1\leqslant k\leqslant K)$ 类中包含的业务个数为 a_k，业务总数 $N=(a_1+a_2+a_3+\ldots+a_K)$。本文主要关注集中式接入网基站的 4 类核心资源，包括 CPU、内存、带宽和 DSP 资源。用 rc_{ij}、rm_{ij}、rbw_{ij}、rdsp_{ij} 分别表示第 $i(1\leqslant i\leqslant K)$ 类中第 j 个业务对 CPU、内存、带宽、DSP 资源的需求，为了表示方便，用矩阵 R_i 表示分类后第 i 类中所有业务资源需求参数矩阵的转置：

$$R_1=\begin{bmatrix} \mathrm{rc}_{11} & \mathrm{rm}_{11} & \mathrm{rbw}_{11} & \mathrm{rdsp}_{11}\\ \mathrm{rc}_{12} & \mathrm{rm}_{12} & \mathrm{rbw}_{12} & \mathrm{rdsp}_{12}\\ & & \vdots & \\ \mathrm{rc}_{1a_1} & \mathrm{rm}_{1a_1} & \mathrm{rbw}_{1a_1} & \mathrm{rdsp}_{1a_1} \end{bmatrix}_{a_1\times 4} \tag{7}$$

$$R_2=\begin{bmatrix} \mathrm{rc}_{21} & \mathrm{rm}_{21} & \mathrm{rbw}_{21} & \mathrm{rdsp}_{21}\\ \mathrm{rc}_{22} & \mathrm{rm}_{22} & \mathrm{rbw}_{22} & \mathrm{rdsp}_{22}\\ & & \vdots & \\ \mathrm{rc}_{2a_2} & \mathrm{rm}_{2a_2} & \mathrm{rbw}_{2a_2} & \mathrm{rdsp}_{2a_2} \end{bmatrix}_{a_2\times 4} \tag{8}$$

$$R_k=\begin{bmatrix} \mathrm{rc}_{k1} & \mathrm{rm}_{k1} & \mathrm{rbw}_{k1} & \mathrm{rdsp}_{k1}\\ \mathrm{rc}_{k2} & \mathrm{rm}_{k2} & \mathrm{rbw}_{k2} & \mathrm{rdsp}_{k2}\\ & & \vdots & \\ \mathrm{rc}_{ka_k} & \mathrm{rm}_{ka_k} & \mathrm{rbw}_{ka_k} & \mathrm{rdsp}_{ka_k} \end{bmatrix}_{a_k\times 4} \tag{9}$$

其中，矩阵 R_i 表示的是第 i 类业务中 a_i 个业务所需的 4 项基站处理资源参数。R_i 的第一行参数表示某个业务所需要的各类计算资源参数大小。

步骤 4 确定每类业务的具体业务类型 type_i

Fisher 最优 K 分割基于业务处理资源需求参数大小的相似性把业务分成 K 类，在第 2.1.2 节中需要根据业务的各类资源需求偏重情况来完成资源分配，所以要确定某类业务的具体业务类型，包括 CPU 型、内存型、带宽型、DSP 型和平衡型，具体方法如下。首先将 R_i 第一行业务作为 Fisher 分割算法的分割点，表明它的资源参数与其他类业务差异很大，取出各类业务的第一行业务所需资源参数依次放入矩阵 Q，利用式（2）进行归一化，使得数值大小在 $[0,1]$ 之间。依据式（10）~式（13）求得该业务某类资源占所有资源总量的比重如下：

$$p(\mathrm{CPU})=\frac{\mathrm{rc}_{k1}}{\mathrm{rc}_{k1}+\mathrm{rm}_{k1}+\mathrm{rbw}_{k1}+\mathrm{rdsp}_{k1}} \tag{10}$$

$$p(\text{内存})=\frac{\mathrm{rm}_{k1}}{\mathrm{rc}_{k1}+\mathrm{rm}_{k1}+\mathrm{rbw}_{k1}+\mathrm{rdsp}_{k1}} \tag{11}$$

$$p(\text{带宽})=\frac{\mathrm{rbw}_{k1}}{\mathrm{rc}_{k1}+\mathrm{rm}_{k1}+\mathrm{rbw}_{k1}+\mathrm{rdsp}_{k1}} \tag{12}$$

$$p(\mathrm{DSP})=\frac{\mathrm{rdsp}_{k1}}{\mathrm{rc}_{k1}+\mathrm{rm}_{k1}+\mathrm{rbw}_{k1}+\mathrm{rdsp}_{k1}} \tag{13}$$

求上述 4 个参数的最大值 $\max(p(\mathrm{CPU}),p(\text{内存}),p(\text{带宽}),p(\mathrm{DSP}))$，所对应的资源类型即代表该业务类型 type_i。当计算 $R_i(1\leqslant i\leqslant K)$ 第一行业务出现两类或三类计算资源所占比重相同的情况时，则继续计算 $R_i(1\leqslant i\leqslant K)$ 第二行业务，直到出现 $(p(\mathrm{CPU}),p(\text{内存}),p(\text{带宽}),p(\mathrm{DSP}))$ 4 个值不等时求得业务类型 type_i。业务类型 type_i 用式（14）来表示：

$$\mathrm{type}_i=\begin{cases} \text{CPU型},\ i=1\\ \text{内存型},\ i=2\\ \text{宽带型},\ i=3\\ \text{DSP型},\ i=4 \end{cases} \tag{14}$$

当 $(p(\text{CPU}), p(\text{内存}), p(\text{宽带}), p(\text{DSP}))$ 的数值大小都在 $(0.25 \pm \Delta)$ 附近，将业务类型定义为 type_i = 平衡型，i=5，Δ 的值大小根据对资源的重要程度来确定，一般取值范围为 0.01~0.1。

基于以上分析可以得出，在某个周期内业务经 Fisher 最优 K 分割被分成 K 类，根据式（10）~式（13）确定的业务类型最多是 5 类（CPU 型、内存型、带宽型、DSP 型和平衡型）。考虑到本文设计的业务类型为 5 类，而最优 K 分割当 $K=1,2,3,4$ 时，不能满足将业务分为 5 类的要求，另外，虽然 K 值越大，分类的准确度越高，但其代价是更高的计算量，当 $K \geqslant 6$ 时，K 值每增加 1，在业务总量为 N 时，需要增加 N 次加法和（N-1）次比较，综合而言，本文的 K 取值为 5。

2.1.2　资源分配均衡模型及分析

根据用户业务对基站处理资源（CPU、内存、带宽、DSP 资源）需求的差异，在物理服务器上建立相应的处理实体（processing entity，PE）处理用户业务。为了实现业务处理的隔离和安全性，规定一个处理实体对应一个业务处理。传统的思路会根据处理资源（以 CPU、内存资源为例）需求来分配资源，如图 1(a)所示，其中大矩形表示一个物理服务器的 CPU、内存资源的总量，图 1(a)的服务器支撑了两个对内存需求较高的用户业务，分别为 PE1 和 PE2。图 1(a)中深色区域表示物理服务器所剩的可用处理资源，由图 1 可见，此物理服务器还剩余大量的 CPU 资源，但内存资源已经不足，难以建立一个新的 PE。如果有第 3 个用户业务请求，在建立对应的处理实体时需要再开启另一个物理服务器，这样就造成图 1(a)的物理服务器 CPU 资源的浪费，出现了资源分配不均衡的问题。

针对该问题，需要将分属不同类的处理资源需求互补的 PE 配置在同一个物理服务器，充分利用 CPU、内存等资源，尽量减少物理服务器开启的数量，如图 2(a)的物理服务器所示。本文以提高资源利用率为目标，根据 Fisher 分割法将某一周期内的业务根据对不同类计算资源需求不同分类，考虑到 5 种业务类型，把分类数 K 设置为 5。定义资源均衡利用率指标体现多类计算资源的综合利用情况，在资源分配过程中充分考虑了资源总量的约束和分配基站处理资源的顺序，如图 2 所示，用资源均衡利用率和各类资源占用率两个指标来分配基站处理资源。相对于经典算法首次适应算法（first fit algorithm，FFA），FFA 的主要思想是按顺序查找物理服务器，将 PE 直接部署在满足基站处理资源要求的物理服务器上，FFA 下的物理服务器的资源使用情况如图 1 所示，本文的算法更有优势。

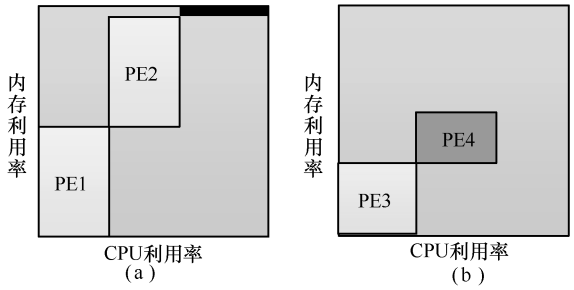


图 1　FFA 下物理服务器的资源利用情况

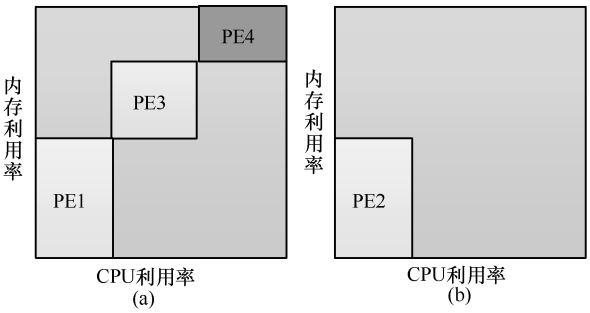


图 2　SPRAA 下物理服务器的资源利用情况

本文综合考虑 CPU、内存、带宽、DSP 资源，设每个物理服务器的资源量相同，其中 CPU 总量为 SC，内存总量为 SM，带宽总量为 SBW。DSP 资源是以 DSP 芯片的形式呈现，用芯片数量 SD=4 来表示 DSP 资源总量。系统中共配置 M 个服务器，用一个矩阵 S 来描述物理服务器的资源配置情况：

$$S = M\begin{bmatrix}\text{SC} & \text{SM} & \text{SBW} & \text{SD}\end{bmatrix} \quad (15)$$

本文通过以下两个指标来衡量业务需求组合的优越性。

（1）定义单个物理服务器 l 的资源均衡利用率表示为：

$$\mu_l = a u_1^{\text{cpu}} + \beta u_1^{\text{mem}} + \gamma u_l^{\text{bw}} + \chi u_l^{\text{dsp}} \tag{16}$$

其中，a、β、γ、χ 表示资源的重要程度，且满足 $\alpha+\beta+\gamma+\chi=1$。

假设物理服务器 l 上加载了 L 个处理实体（PE），则物理服务器 l 的资源均衡利用率为：

$$\mu_l = \alpha \frac{\sum_{i=1}^{L} \text{rc}_i}{\text{SC}} + \beta \frac{\sum_{i=1}^{L} \text{rm}_i}{\text{SM}} + \gamma \frac{\sum_{i=1}^{L} \text{rbw}_i}{\text{SBW}} + \chi \frac{\sum_{i=1}^{L} \text{rdsp}_i}{\text{SD}} \tag{17}$$

其中，$\sum_{i=1}^{L} \text{rc}_i$ 表示 L 个处理实体所需的 CPU 总和，$\sum_{i=1}^{L} \text{rm}_i$ 表示 L 个处理实体所需的内存资源总和，$\sum_{i=1}^{L} \text{rbw}_i$ 表示 L 个处理实体所需的带宽资源总和，$\sum_{i=1}^{L} \text{rdsp}_i$ 表示 L 个处理实体所需的 DSP 资源总和。μ_l 越大，资源利用越均衡。当处理实体选择建立的物理服务器时，通过计算：若将此处理实体加载到已开启的物理服务器的资源均衡利用率 μ_l 的大小，将 μ_l 最大值的物理服务器作为拟选定的物理服务器。在 α、β、γ、χ 分别为 0.4、0.2、0.2、0.2 的前提下，计算 μ_l 可能会出现相等的情况，所以定义各类计算资源占用情况作为分配资源的另一个指标，如（2）所示。

（2）物理服务器 l（$1 \leqslant l \leqslant M$）的资源占用情况

用四元组 $u_l = (u_l^{\text{cpu}}, u_l^{\text{mem}}, u_l^{\text{bw}}, u_l^{\text{dsp}})$ 表示，其中 u_l^{cpu}、u_l^{mem}、u_l^{bw}、u_l^{dsp} 分别表示物理服务器 l 的 CPU、内存、带宽、DSP 资源的利用率。这个四元组在资源均衡利用率相等的情况下，更加清晰地表示出每类计算资源的占用情况。用这个参数衡量业务需求组合的优越性有一定的指导意义。

基于以上分析，如给内存型业务对应的处理实体 i 分配处理资源时，依次计算出将内存型的处理实体（PE）加载在已开启服务器的资源均衡利用率，选择资源均衡利用率高的物理服务器作为拟选定的物理服务器。若出现相同的资源均衡利用率的情况，则利用四元组 $u_l = (u_l^{\text{cpu}}, u_l^{\text{mem}}, u_l^{\text{bw}}, u_l^{\text{dsp}})$ 选择在内存资源占用率低的物理服务器上作为拟选定的物理服务器。将此处理实体 PE i 加载在拟选定的物理服务器 l 的资源总量约束条件为：

$$\text{rc}_i \leqslant \text{SC}_l^{\text{res}} \tag{18}$$

$$\text{rm}_i \leqslant \text{SM}_l^{\text{res}} \tag{19}$$

$$\text{rbw}_i \leqslant \text{SBW}_l^{\text{res}} \tag{20}$$

$$\text{rdsp}_i \leqslant \text{SD}_l^{\text{res}} \tag{21}$$

其中，rc_i、rm_i、rbw_i、rdsp_i 分别表示处理实体 i 所需要的 CPU、内存、带宽和 DSP 资源。SC_l^{res}、SM_l^{res}、$\text{SBW}_l^{\text{res}}$、$\text{SD}_l^{\text{res}}$ 表示物理服务器 l 剩余的 CPU、内存、带宽和 DSP 资源。

在某个周期内的资源分配完成时，假设资源池中开启了 M_{phy} 个物理服务器，则整个资源池的资源利用率 u_{sum} 为：

$$u_{\text{sum}} = \frac{\sum_{l=1}^{M_{\text{phy}}} \mu_l}{M_{\text{phy}}} \tag{22}$$

在某个周期内，本文根据用户业务对各类计算资源需求参数的不同，将业务请求分为 K=5 类，业务类型由式（10）~式（14）来确定。在资源分配过程中尽量将不同业务需求的处理实体分配在同一个物理服务器，避免物理服务器出现严重的资源分配不均衡现象。在集中式接入网架构下基于用户业务的资源分配步骤如下。

步骤 1 采用 Fisher 最优分割法对用户业务进行分类，建立业务处理资源需求参数矩阵 $R_i(1 \leqslant k \leqslant K)$。

步骤 2 初始化物理服务器 $l(1 \leqslant l \leqslant M)$ 的各类资源利用率 $u_l = (0,0,0,0)$ 初始时刻每个物理服

务器上加载的 PE 个数均为 0。初始化 $i=1$，$j=1$。

步骤 3　根据处理资源需求 $R_i(1\leqslant k\leqslant K)$ 确定第 j 个业务的 PE 建立的物理服务器。考虑到不同类型的业务数量不等的情况，首先对 j 进行判断，若 $j>a_i$，则转至步骤 7；否则执行步骤 4。

步骤 4　若给内存型的处理实体分配资源时，依次计算出将内存型的 PE 加载在已开启服务器的资源均衡利用率，选择资源均衡利用率高的物理服务器作为拟选定的物理服务器。若出现相同的资源均衡利用率的情况，则利用四元组 $u_l=(u_l^{\text{cpu}},u_l^{\text{mem}},u_l^{\text{bw}},u_l^{\text{dsp}})$ 选择在内存资源占用率低的物理服务器上作为拟选定的物理服务器；检查拟选定物理服务器是否满足 PE 建立所需的 CPU、内存、带宽和 DSP 资源大小，即 $\text{rc}_i\leqslant\text{SC}_l^{\text{res}}$、$\text{rm}_i\leqslant\text{SM}_l^{\text{res}}$、$\text{rbw}_i\leqslant\text{SBW}_l^{\text{res}}$、$\text{rdsp}_i\leqslant\text{SD}_l^{\text{res}}$。若能满足，将PE建立在此物理服务器上并转至步骤6；若不满足，执行步骤 5。

步骤 5　开启一个新的物理服务器，将 PE 建立在此物理服务器上，转至步骤 6。

步骤 6　记录 PE 的处理资源参数（CPU、内存、带宽和 DSP 资源），为计算整体资源利用率提供数据；更新物理服务器剩余的处理资源、各类资源的利用率 u_l 和整体资源利用率 u_{sum}。

步骤 7　改变参数：$i=i+1$；当 $i>K$ 时，$j=j+1$，$i=1$，返回步骤 3。

2.2　集中式接入网架构下的基站处理资源管理架构

基于第 2.1 节的基站处理资源分配算法，本文给出了集中式接入网中的基站处理资源管理架构，如图 3 所示，完成对基站处理资源的统一管理和按需分配。业务监测模块主要监测用户业务请求的到来，将到达的业务传到业务分类模块进行处理。业务分类模块采用分割算法完成对用户业务的分类，交给资源分配模块为业务建立 PE。资源分配模块主要根据资源维护模块提供的各类计算资源利用率和剩余资源参数来进行处理资源的分配，建立业务的 PE。监控设备主要用来收集系统中各类资源的利用率和物理服务器上剩余的各类资源参数，通过性能分析为处理资源管理中心调整资源分配算法提供依据，同时监测硬件资

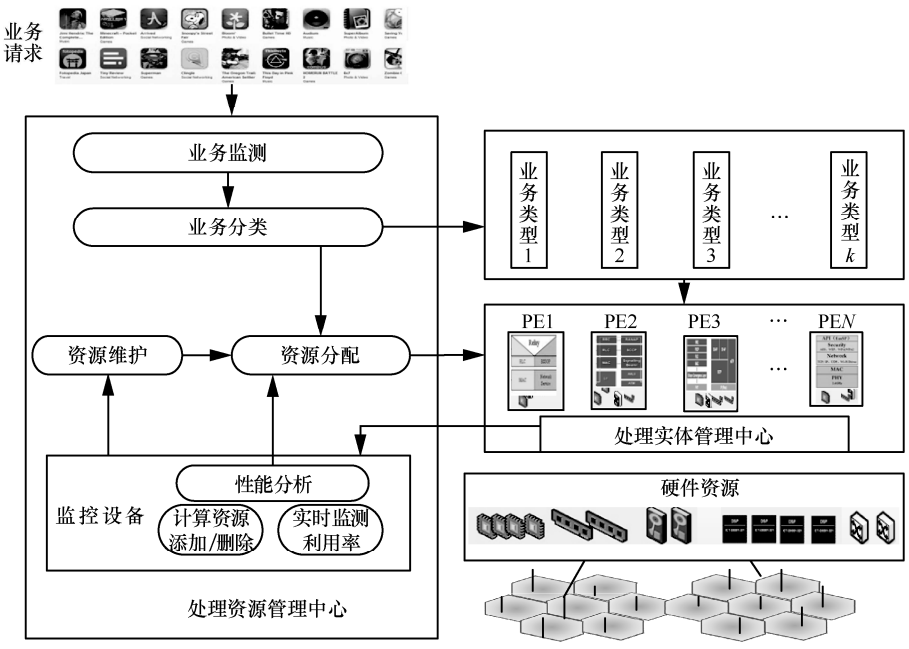


图 3　集中式接入网架构下的基站处理资源管理架构

源的状态变化（如处理资源的添加或删除等），把这些信息上报给信息维护模块，维持处理资源的实时信息。处理实体管理中心记录了每个物理服务器上加载的 PE 的各类资源参数大小，把这份数据上报给监控设备，为计算整体资源利用率 u_{sum} 提供数据。

3 仿真分析

本文通过在 MATLAB2012a 仿真验证所提算法的性能。综合考虑了物理服务器中的 CPU、内存、网络带宽、DSP 资源，将基站处理实体分为 K=5 类（CPU 型、内存型、带宽型、DSP 型和平衡型），并按上文给的资源分配算法进行处理资源的分配。为了验证算法的有效性，本文与经典算法 FFA 做比较。

在本文的仿真场景中，假设物理服务器的个数能够满足业务需求，在实验中更关注物理服务器开启的个数。假定业务的到达服从泊松过程（λ=5），即平均每秒到达业务数为 5。参照参考文献[20]，在分类数 K 确定为 5 的情况下，当数据样本数为 50~120 时，Fisher 算法的分割效果达到最佳。为了获得这些样本数，考虑到平均每秒到达业务数为 5，相应的时间区间应以 10~24 s 为宜，以便达到较好的 Fisher 分割效果；同时考虑到业务处理的及时性，处理的时间区间应尽量小，所以本文以 10 s 为一个周期，对 10 s 内到达的业务进行分类和建立处理业务的 PE，在第 10 s 时刻计算总体资源利用率 u_{sum}。

在建立业务对应的处理实体时，资源开销主要包括两部分：

（1）业务本身对各类资源的需求，即 $[\text{rc}_i\ \text{rm}_i\ \text{rbw}_i\ \text{rdsp}_i]$；

（2）当多个处理实体加载到同一个物理服务器时，这些处理实体对应的数据处理和控制相关进程会在物理服务器中调度执行，这些调度本身也会产生处理资源开销，比如多个处理实体进程切换时的进程激活，同一个处理实体内部进程间调度，这些都会造成 CPU 资源的开销。

本文主要考虑到业务本身对各类资源的需求，为了表征（2）对 CPU 资源的消耗，把 CPU 资源的重要参数设置得大一点。在仿真中，把 CPU 资源的重要参数设置为 0.4，内存、带宽和 DSP 资源统一设置为 0.2。

仿真参数设置见表 1。

表 1 仿真参数设置

参数说明	参数值
物理服务器核数/个	8
物理服务器内存/GB	8
网络带宽/（$\text{Mbit}\cdot\text{s}^{-1}$）	100
DSP 芯片数量/个	4
业务类型（处理实体类型）K	5
某个周期内到达的业务总数	N
开启的服务器个数	M
资源重要程度参数 a、β、γ、χ	0.4、0.2、0.2、0.2
变化量Δ	0.04

图 4 对比了所提算法 SPRAA 和 FFA 的整个资源池物理服务器资源利用率 u_{sum} 随时间变化的情况。可以看出，随着时间的增加，用 FFA 分配资源，资源利用率会起伏不定，主要是 FFA 按顺序查找物理服务器，找到符合条件的就将处理实体加载到物理服务器上，而用所提算法 SPRAA 会考虑到物理服务器的各类资源利用率，将处理资源需求互补的处理实体组合到一起，随着时间增加，到达的业务数增加，能够找到更多处理资源需求互补的业务进行组合，所以资源利用率会逐渐增加。当时间到达一定数量后，SPRAA 的资源利用率会明显高于 FFA，主要是因为 FFA 出现了第 2.1.2 节中图 1 的资源分配不均衡的情况，不能充利用分物理服务器的各类资源；而用 SPRAA 分配资源，可以使得物理服务器上加载的 PE 互补，最大化地利用基站处理资源，SPRAA 优势更明显。在实际中，由于不同类的用户业务对资源需

求参数不可能出现完全互补的情况，使得物理服务器的资源利用率不能达到 100%，因此当时间到达一定数量后，物理服务器的资源利用率的增加趋于平缓，可达到的利用率在 70%左右。

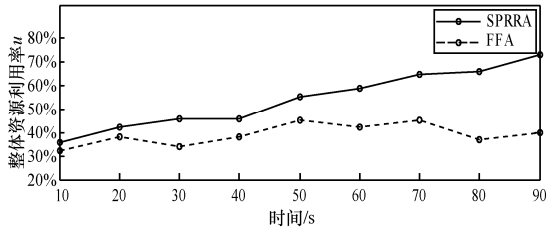


图 4　物理服务器资源利用率对比

图 5 对比了所提算法和 FFA 的物理服务器开启数目。可以看出，本文提出的算法能减少开启物理服务器的个数，主要因为在资源分配过程中尽量将不同业务需求的处理实体配置在同一个物理服务器上，最大限度地利用 CPU、内存、网络带宽和 DSP 资源。

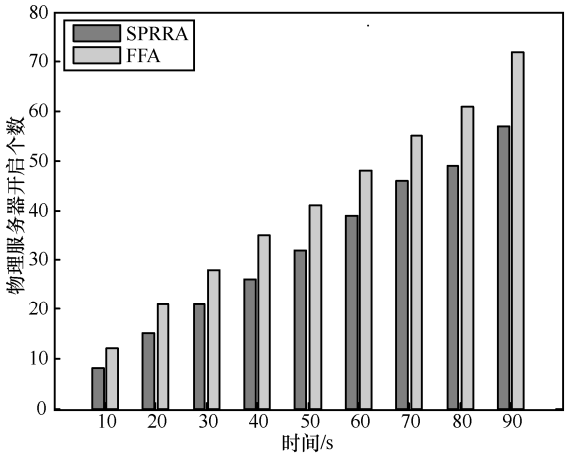


图 5　物理服务器开启数目对比

4　结束语

本文在分析了集中式架构下资源分配的研究现状后，考虑到用户业务类型对不同类型计算资源的需求，提出了一种基于业务类型的集中式接入网的基站处理资源分配算法。在资源分配过程中，将不同的处理资源需求互补的业务组合到一起，解决了物理服务器的资源分配不均衡的问题，仿真表明所提算法减少了开启物理服务器的个数，同时提高了物理服务器的资源利用率。在下一步的研究中，若用户某种业务突然增加，可能会造成 PE 的迁移，采用的迁移算法仍需要进一步研究。

参考文献：

[1] LIU L, ZHOU Y Q, TIAN L, et al. Load aware joint CoMP clustering and inter-cell resource scheduling in heterogeneous ultra dense cellular networks[J]. IEEE Transactions on Vehicular Technology, Early Access, 2017, 99(11): 1-14.

[2] ZHOU Y Q, LIU H, PAN Z, et al. Cooperative multicast with location aware distributed mobile relay selection: performance analysis and optimized design[J]. IEEE Transactions on Vehicular Technology, 2017, 66(3): 8291-8302.

[3] China Mobile. C-RAN: the road towards green RAN[R]. 2015.

[4] QIAN M, WANG Y, ZHOUY, et al. A super base station based centralized network architecture for 5G mobile communication systems [J]. Digital Communications & Networks, 2015, 1(2): 152-159.

[5] HUANG L, ZHOU Y. Coverage optimization for femtocell clusters using particle swarm optimization[C]//2012 IEEE International Conference on Communications, June 10-15, 2012, Ottawa, ON, Canada. Piscataway: IEEE Press, 2012: 612-615.

[6] 田霖, 翟国伟, 黄亮, 等. 基于集中式接入网架构的异构无线网络资源管理技术研究[J]. 电信科学, 2013, 29(6): 25-27.
TIAN L, ZHAI G W, HUANG L, et al. Research on key technologies of heterogeneous wireless network resource management based on centralized radio access network architecture[J]. Telecommunications Science, 2013, 29(6): 25-27.

[7] ZHAI G W, TIAN L, ZHOU Y Q, et al. Load diversity based optimal processing resource allocation for super base stations in centralized radio access networks[J]. Science China Information Sciences, 2014, 57(4): 1-12.

[8] MELL P, GRANCE T. The NIST definition of cloud computing(draft)[R]. 2013.

[9] 殷波, 张云勇, 房秉毅, 等. 面向成本优化的云计算资源分配方法研究[J]. 电信科学, 2014, 30(11): 22-26, 32.
YIN B, ZHANG Y Y, FANG B Y, et al. Cloud resource allocation method based on elastic resource adjustment[J]. Telecommunications Science, 2014, 30(11): 22-26, 32.

[10] 吴清烈, 郭昱, 武忠. 云计算服务与大规模定制模式应用[J]. 电信科学, 2010, 26(9): 74-78.
WU Q L, GUO Y, WU Z. Study on cloud computing services and mass customization applications[J]. Telecommunications Science, 2010, 26(9): 74-78.

[11] 徐骁勇, 柯涛, 刘梦娟, 等. 面向云平台的资源分配策略研

究[J]. 计算机应用, 2013, 33(20): 299-307.
XU X Y, KE T, LIU M J, et al. Research on resource allocation strategies in cloud computing[J]. Journal of Computer Applications, 2013, 33(20): 299-307.

[12] 陈小娇, 陈世平, 方芳. 云计算中虚拟机资源分配算法[J]. 计算机应用研究, 2014, 31(9): 2585-2616.
CHEN X J, CHEN S P, FANG F. Research on virtual machine resource allocation algorithm in cloud computing[J]. Journal of Computer Applications Research, 2014, 31(9): 2585-2616.

[13] LU J. Improved genetic algorithm-based resource scheduling strategy in cloud computing[C]//2016 International Conference on Smart City and Systems Engineering, Nov 25-26, 2016, Changsha, China. Piscataway: IEEE Press, 2016: 230-234.

[14] WANG S, GU H, WU G. A new approach to multi-objective virtual machine placement in virtualized data center[C]//IEEE Eighth International Conference on Networking, Architecture and Storage, July 17-19, 2013, Xi'an, China. Piscataway: IEEE Press, 2013: 331-335.

[15] CHEN N S, FANG X P. A cloud computing resource scheduling scheme based on estimation of distribution algorithm[C]//The 2014 2nd International Conference on Systems and Information, Nov 15-17, 2014, Shanghai, China. Piscataway: IEEE Press, 2014: 304-308.

[16] GHRIBI C, HADJI M, ZEGHLACHE D. Energy efficient VM scheduling for cloud data centers: exact allocation and migration algorithms[C]//2013 IEEE/ACM 13th International Symposium on Cluster, Cloud, and Grid Computing, May 13-16, 2013, Delft, the Netherlands. Piscataway: IEEE Press, 2013: 671-678.

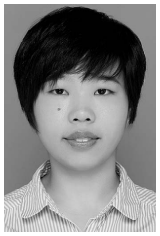
[17] IMT-2020(5G)推进组. 5G 概念白皮书[R]. 2015.
IMT-2020(5G) Propulsion Group. 5G white paper [R]. 2015.

[18] 翟国伟. 超级基站架构下协议处理资源分配和管理方法的研究[D]. 北京: 中国科学院大学, 2016.
ZHAI G W. Research on protocol processing resource allocation and management in super BS architecture [D]. Beijing: University of Chinese Academy of Sciences, 2016.

[19] 王庆扬, 谢沛荣, 熊尚坤, 等. 5G 关键技术与标准综述[J]. 电信科学, 2017, 33(11): 112-122.
WANG Q Y, XIE P R, XIONG S K, et al. Key technology and standardization progress for 5G[J]. Telecommunications Science, 2017, 33(11): 112-122.

[20] 武琳琳. 基于 Fisher 最优分割法的聚类分析应用[D]. 郑州: 郑州大学, 2013.
WU L L. Cluster analysis based on the method of Fisher optimal division[D]. Zhengzhou: Zhengzhou University, 2013.

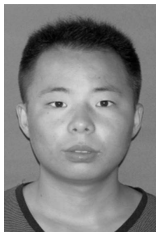
[作者简介]



张新苹（1992−），女，重庆邮电大学硕士生，主要研究方向为绿色无线通信系统、5G。

王园园（1986−），女，博士，中国科学院计算技术研究所助理研究员，主要研究方向为绿色无线通信系统、5G、无线接入网络虚拟化等。

田霖（1980−），女，博士，中国科学院计算技术研究所副研究员，主要研究方向为绿色无线通信系统与无线资源管理技术。



郝树良（1991−），男，重庆邮电大学硕士生，主要研究方向为 5G 新型多址接入技术。

研究与开发

基于有限反馈的毫米波 MIMO 系统的混合预编码方法

尤若楠，潘鹏，张丹，王海泉
（杭州电子科技大学通信工程学院，浙江 杭州 310016）

摘　要：针对发送端未知信道状态信息下毫米波 MIMO 系统的混合预编码问题，提出了一种基于有限反馈的混合预编码算法。该算法首先将模拟及数字预编码的混合优化问题简化为模拟预编码和数字预编码独立优化问题，模拟预编码将依据该独立优化函数在所设计的模拟预编码码本内选取最佳的预编码矩阵；然后，根据已获得的模拟预编码矩阵，并基于最小二乘法获得数字预编码矩阵，在数字预编码码本内选择与其距离最近的码字作为反馈数字预编码矩阵；最后，接收机将模拟和数字预编码矩阵的索引值反馈回发射机。仿真结果显示，所提算法能够在复杂度和性能上实现较好的均衡。

关键词：毫米波；多输入多输出；预编码；码本设计；有限反馈
中图分类号：TN928
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018163

Hybrid precoding method for mmWave MIMO systems based on limited feedback

YOU Ruonan, PAN Peng, ZHANG Dan, WANG Haiquan
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310016, China

Abstract: A hybrid precoding algorithm for millimeter wave (mmWave) MIMO systems without knowledge of channel state information at the transmitter was proposed based on the limited feedback. Specifically, the joint optimization problem of analog precoding and digital precoding was firstly separated into two independent optimization problems, which correspond to the analog precoding and digital precoding respectively. Based on this analog precoding codebook, the best precoding matrix was selected by analog precoding in the designed analog precoding codebook. Secondly, according to the obtained analog precoding matrix, the digital precoding matrix was obtained through the least square method, and then, the codeword closest to it in the random vector quantization codebook was selected as the feedback digital precoding matrix. Finally, the receiver feeds the index values of the analog and digital precoding matrices back to the transmitter. Simulation results show that the proposed algorithm can achieve better balance between complexity and performance.

Key words: millimeter wave, multiple input multiple output, precoding, codebook design, limited feedback

收稿日期：2018−01−03；修回日期：2018−04−19
基金项目：国家自然科学基金资助项目（No.61401130）
Foundation Item: The National Natural Science Foundation of China (No.61401130)

1 引言

以智能手机为代表的移动终端已经成为人们日常生活中必不可少的设备，并使得以多媒体为代表的新兴业务逐渐代替了原有以语音/短信为主体的传统业务，由此对移动通信网络的传输速率和服务质量提出了更高的要求。然而，有限的频谱资源始终是限制高速数据传输的关键因素，尽管现有移动通信系统通过多输入多输出（MIMO）、正交频分复用和多小区协作等技术已经极大地提升了频谱效率。随着5G时代的来临，剩余的低频段频谱资源已经不能满足所提出的10 Gbit/s峰值速率的需求，因此未来的5G系统需要在更高频段上寻找新的可用资源。基于此，工作于26.5 GHz以上的毫米波（millimeter wave，mmWave）通信技术已成为5G最有希望的候选技术之一，这是因为其相比于传统低频微波频段具有更宽的带宽，且很多目前尚未授权使用[1-5]。但是，考虑到毫米波信号的波长较短，与目前得到广泛使用的6 GHz以下的微波信号相比具有更严重的路径损耗；同时，也正得益于毫米波较短的波长，使其在一定的尺寸内封装大量的天线元件成为可能，从而可利用大规模天线阵列产生的波束成形增益来弥补所造成的路径损耗。因此，在实际的毫米波通信系统中，大规模天线阵列的应用必不可少。

在传统的低频多天线系统中，波束成形和预编码一般在基带通过数字信号处理单元进行处理，这就需要为每根天线配备专门的射频（radio frequency，RF）链路[6]。然而，由于毫米波天线阵列的天线数目往往较多，且毫米波电路的硬件成本和功率损耗均较高，这使得目前基于全数字预编码/解码的收发器架构难以应用于毫米波频段[10]。因此，模数混合预编码/解码架构被提出[6-8,11-13]，并受到了广泛的关注和研究。在模数混合预编码架构中，发送端首先将并行的多个数据流经基带数字预编码后再映射到若干个RF链路上，然后通过模拟的相移网络实现模拟预编码，最后经由阵列天线实现信号发送。在该架构中，射频链路数通常远小于天线数，从而能够极大程度地降低毫米波MIMO系统的硬件复杂度。一般而言，对于模数混合结构，模拟相移网络可以提供波束成形增益（beamforming gain），而数字预编码则可以提供多流或多用户的复用增益（multiplex gain）。在实际系统中，根据RF链路是否与全部天线相连，模数混合架构可以分为全连接架构[6]和部分连接架构[12]，而模拟相移网络通常可以使用移相器[2,6]、开关[9]或者透镜[14]来实现。当使用模拟移相器实现模拟波束成形时，可以通过调整RF链路上信号的相对相位，从而将发送/接收波束引导到预期的方向上。从信号角度来看，模拟相移网络可以看成对基带数字信号进行预编码，只不过此时预编码矩阵的每个元素需满足恒模约束。

通常，为了使毫米波通信系统的频谱效率达到最佳，需要对发送端的数字/模拟预编码矩阵以及接收端的模拟/数字合并矩阵进行联合优化。然而，如参考文献[6]所述，找到在条件约束下的联合优化问题的全局最优解是非常棘手的。即使在传统多用户MIMO（MU-MIMO）系统中，也需要通过交替迭代优化来找到求解和速率最大化的局部最优值[15]。然而，已有的一些混合预处理设计方案[7,8,12]表明，对发射机和接收机进行分开设计也可以获得令人满意的性能，从而降低了系统设计的复杂度，避免了庞杂的迭代优化算法。基于此，本文将主要针对发送端的模拟和数字预编码设计展开研究。此外，所提方法还可以扩展到多小区协作多用户毫米波MIMO或多跳中继MIMO传输系统中[20-24]。对于有基站、中继站和两个用户设备组成的双向多天线中继网络，参考文献[24]基于自适应监听协议提出的迭代算法可以联合优化两个时隙上的用户接收模式时的自适

应权重和第二个时隙中继处的预编码矩阵，然而这种方法会增加额外的开销，即需要知道两个监听信道系数的 CSI，而本文提出的方法恰恰可以降低系统开销和计算复杂度。

本文考虑了采用频分双工（frequency division duplex，FDD）的单用户毫米波 MIMO 系统全连接混合预编码架构，并假设在接收端采用最佳数字解码的情况下，对发送端的模数混合预编码矩阵的优化设计问题展开研究。由于 FDD 下，发送端只能依靠反馈获得预编码信息，因此本文提出了一种基于有限反馈码本的混合预编码方法，首先构造在发送端和接收端两侧都已知的有限模拟预编码码本 F 和数字预编码码本 $\mathcal{W}$；接收端由估计获得的下行信道信息，根据预编码矩阵的优化问题在码本内进行码字搜索，从而获得模拟预编码矩阵和数字预编码矩阵；最后将相应的索引值反馈回发送端。注意到模拟预编码器一般采用模拟移相器来实现，则模拟预编码矩阵的元素会受到恒模约束，所以在构造模拟预编码码本时要考虑这一约束条件。由于数字预编码矩阵没有特别的硬件约束限制，为简单起见，本文选择随机矢量量化码本作为数字预编码码本。本文的主要贡献在于以下两个方面。

- 将模拟预编码和数字预编码的联合优化问题简化为它们各自的优化问题，从而无须进行复杂的迭代优化过程，降低了实现复杂度。此外，该方法有利于双码本结构的实现，从而缩小了码字搜索空间，进一步降低了复杂度。
- 提出了一种应用于毫米波通信的模拟预编码码本构造方法。由于模拟预编码矩阵的元素需要受到恒模约束，因此基于码本内各个码字间最小角度最大化的原理，设计了基于离散傅里叶变换（discrete Fourier transform，DFT）矩阵及其旋转矩阵的码本构造方法。

仿真结果表明，所提出的有限反馈模数混合预编码方法能够在复杂度和系统的频谱效率上实现较好的均衡。

2　系统模型

考虑如图 1 所示的单用户毫米波 MIMO 系统，发送端配备 N_{t} 根发射天线，并向配备有 N_{r} 根接收天线的接收端并行传输 N_{s} 路数据流。与传统多天线系统不同，毫米波系统的 RF 链路将远少于收发天线的数目，如图 1 所示，发送端和接收端的 RF 链路数目分别为 $N_{\mathrm{t}}^{\mathrm{RF}}$（$<N_{\mathrm{t}}$）和 $N_{\mathrm{r}}^{\mathrm{RF}}$（$<N_{\mathrm{r}}$）。为了使毫米波 MIMO 系统能支持并行多数据流通信，所传输的数据流 N_{s} 与收发端的 RF 链路需进一步满足 $N_{\mathrm{s}} \leqslant N_{\mathrm{t}}^{\mathrm{RF}} \leqslant N_{\mathrm{t}}$ 和 $N_{\mathrm{s}} \leqslant N_{\mathrm{r}}^{\mathrm{RF}} \leqslant N_{\mathrm{r}}$。在发送端，发送符号矢量 $\boldsymbol{s} \in \mathbb{C}^{N_{\mathrm{s}} \times 1}$ 首先进行线性的基带数字预编码，即把发送符号矢量与数字预编码矩阵 $\boldsymbol{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{t}}^{\mathrm{RF}} \times N_{\mathrm{s}}}$ 相乘；然后在 RF 链路内实现数模转换；最后再通过相移网络实现模拟预编码并经阵列天线实现信号的发射。从信号角度来说，以相移网络实现的模拟预编码相当于对信号乘以一个模拟预编码矩阵 $\boldsymbol{F}_{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{t}} \times N_{\mathrm{t}}^{\mathrm{RF}}}$。因此，发送信号最终可以表示为 $\boldsymbol{x}=\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\boldsymbol{s}=\boldsymbol{F}\boldsymbol{s}$，其中，$\boldsymbol{F} \in \mathbb{C}^{N_{\mathrm{t}} \times N_{\mathrm{s}}}$ 是模拟预编码矩阵和数字预编码矩阵的组合。本文假设信息符号具有单位功率，即 $E\left[\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\right]=1/N_{\mathrm{s}}\,\boldsymbol{I}_{N_{\mathrm{s}}}$。由于模拟预编码矩阵一般使用模拟移相器实现，因此模拟预编码矩阵 $\boldsymbol{F}_{\mathrm{RF}}$ 的元素均受恒模限制，即满足 $\left(\boldsymbol{F}_{\mathrm{RF}}^{(\mathrm{i})}\boldsymbol{F}_{\mathrm{RF}}^{(\mathrm{i})\mathrm{H}}\right)_{l,l}=1/N_{\mathrm{t}}$，其中，$\boldsymbol{F}_{\mathrm{RF}}^{(\mathrm{i})}$ 表示模拟预编码矩阵 $\boldsymbol{F}_{\mathrm{RF}}$ 的第 i 列，$(\bullet)_{l,l}$ 则表示矩阵的第 l 个对角线元素。此外，为了满足发送端的总功率限制要求，通过对数字预编码矩阵 $\boldsymbol{F}_{\mathrm{BB}}$ 进行归一化处理，使得 $\left\|\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}^{2}=N_{\mathrm{s}}$。相比于模拟预编码器的恒模约束，基带数字预编码器一般认为没有其他硬件相关的约束限制。

在窄带块衰落信道下[6,8]，若从发送端到接收端的下行链路信道矩阵为 $\boldsymbol{H} \in \mathbb{C}^{N_{\mathrm{r}} \times N_{\mathrm{t}}}$，则接收端的接收信号可表示为：

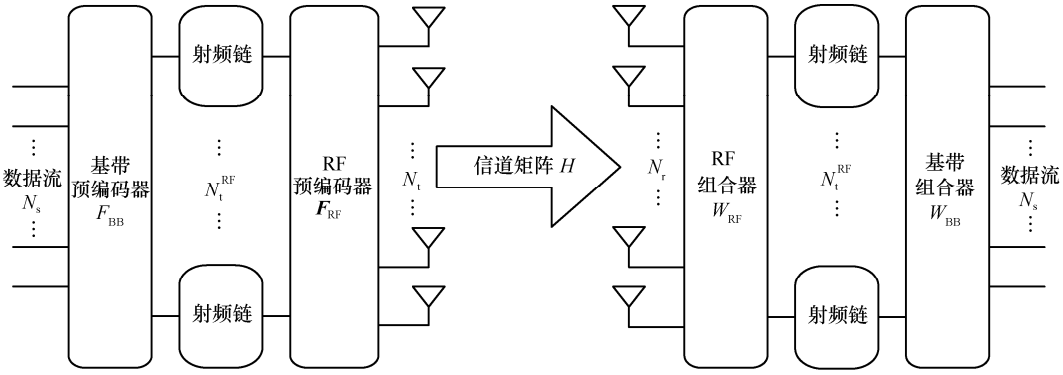


图 1　毫米波单用户系统的混合模拟/数字预编码和组合

$$\boldsymbol{y}=\sqrt{\rho}\boldsymbol{H}\boldsymbol{x}+\boldsymbol{z}=\sqrt{\rho}\boldsymbol{H}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\boldsymbol{s}+\boldsymbol{z} \tag{1}$$

其中，$\boldsymbol{y}\in\mathbb{C}^{N_{\mathrm{r}}\times 1}$ 为接收矢量，ρ 表示平均接收功率，而 $\boldsymbol{z}\in\mathbb{C}^{N_{\mathrm{r}}\times 1}$ 为服从复高斯分布 $\mathcal{CN}\left(0,\sigma^2\boldsymbol{I}_{N_{\mathrm{r}}}\right)$ 的白噪声矢量。

与发送端类似，接收端一般也将采用模数混合方式对接收信号进行处理和检测。令其基带数字接收处理矩阵和模拟接收合并矩阵分别为 $\boldsymbol{W}_{\mathrm{BB}}\in\mathbb{C}^{N_{\mathrm{r}}^{\mathrm{RF}}\times N_{\mathrm{s}}}$ 和 $\boldsymbol{W}_{\mathrm{RF}}\in\mathbb{C}^{N_{\mathrm{t}}\times N_{\mathrm{r}}^{\mathrm{RF}}}$，并令 $\boldsymbol{W}=\boldsymbol{W}_{\mathrm{RF}}\boldsymbol{W}_{\mathrm{BB}}$ 为处理接收信号 $\boldsymbol{y}$ 的接收处理矩阵，则处理后的关于发送数据的估计矢量为：

$$\hat{\boldsymbol{s}}=\boldsymbol{W}_{\mathrm{BB}}^{\mathrm{H}}\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{H}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\boldsymbol{s}+\boldsymbol{W}_{\mathrm{BB}}^{\mathrm{H}}\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{z} \tag{2}$$

由于参考文献[6]已证明发送端和接收端可以分别进行优化设计，且能取得接近收发端联合优化设计的性能；因此，本文主要关注发送端的预处理算法设计。同时为简单起见，假设接收端采用最佳的数字译码，此时系统的频谱效率为：

$$\boldsymbol{R}=\mathrm{lb}\left(\left|\boldsymbol{I}_{N_{\mathrm{r}}}+\boldsymbol{H}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\times\boldsymbol{F}_{\mathrm{BB}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\right|\right) \tag{3}$$

将此作为设计数字预编码矩阵 $\boldsymbol{F}_{\mathrm{BB}}$ 和模拟预编码矩阵 $\boldsymbol{F}_{\mathrm{RF}}$ 的性能指标。

考虑到毫米波信号的传播特性，毫米波系统一般配置较大的天线阵列，并采用波束成形技术，从而将发射功率引导到预期方向，以弥补较高的路径损耗；此外，由于毫米波的波长较短，使得其具有稀疏的散射特性。因此，传统 MIMO 系统的信道矩阵模型将不再适用于毫米波系统，所以本文采用基于扩展的 Saleh-Valenzuela 几何信道模型的窄带信道表示。则信道矩阵 $\boldsymbol{H}$ 可以表示为：

$$\boldsymbol{H}=\sqrt{\frac{N_{\mathrm{r}}N_{\mathrm{t}}}{N_{\mathrm{cl}}N_{\mathrm{ray}}}}\sum_{i=1}^{N_{\mathrm{cl}}}\sum_{l=1}^{N_{\mathrm{ray}}}\alpha_{i,l}\boldsymbol{a}_{\mathrm{r}}\left(\phi_{i,l}^{\mathrm{r}},\theta_{i,l}^{\mathrm{r}}\right)\boldsymbol{a}_{\mathrm{t}}^{\mathrm{H}}\left(\phi_{i,l}^{\mathrm{t}},\theta_{i,l}^{\mathrm{t}}\right) \tag{4}$$

其中，N_{cl} 和 N_{ray} 分别表示簇的数目和每个簇的路径数；$\alpha_{i,l}$ 表示第 i 个簇中的第 l 条路径的信道增益，服从复高斯分布 $\alpha_{i,l}\sim\mathcal{CN}\left(0,\ \sigma_{\alpha,i}^2\right)$，其方差 $\sigma_{\alpha,i}^2$ 表示第 i 个簇的平均功率，并且满足 $\sum_{i=1}^{N_{\mathrm{cl}}}\sigma_{\alpha,i}^2=\beta$，此处定义 $\beta=\sqrt{N_{\mathrm{t}}N_r/N_{\mathrm{cl}}N_{\mathrm{ray}}}$ 为标准化因子，使得 $\mathbb{E}\left\{\left\|\boldsymbol{H}\right\|_{\mathrm{F}}^2\right\}=N_{\mathrm{t}}N_{\mathrm{r}}$。此外，在式（4）中，$\left(\phi_{il}^{\mathrm{r}},\theta_{il}^{\mathrm{r}}\right)$ 是第 i 个簇中的第 l 条路径在水平（方位）和垂直（俯仰）方向上的到达角（angle of arrival，AOA）；$\left(\phi_{il}^{\mathrm{t}},\theta_{il}^{\mathrm{t}}\right)$ 则是同一条路径在水平（方位）和垂直（俯仰）方向上的出发角（angle of departure，AOD）。基于上述角度，向量 $\boldsymbol{a}_{\mathrm{r}}\left(\phi_{il}^{\mathrm{r}},\theta_{il}^{\mathrm{r}}\right)$ 和 $\boldsymbol{a}_{\mathrm{t}}\left(\phi_{il}^{\mathrm{t}},\theta_{il}^{\mathrm{t}}\right)$ 分别表示为归一化的接收阵列响应矢量和发送阵列响应矢量；且由它们组成的矩阵 $\boldsymbol{A}_{\mathrm{r}}=\left[\boldsymbol{a}_{\mathrm{r}}\left(\phi_{1,1}^{\mathrm{r}},\ \theta_{1,1}^{\mathrm{r}}\right),\cdots,\ \boldsymbol{a}_{\mathrm{r}}\left(\phi_{N_{\mathrm{cl}},N_{\mathrm{ray}}}^{\mathrm{r}},\ \theta_{N_{\mathrm{cl}},N_{\mathrm{ray}}}^{\mathrm{r}}\right)\right]$ 和 $\boldsymbol{A}_{\mathrm{t}}=\left[\boldsymbol{a}_{\mathrm{t}}\left(\phi_{1,1}^{\mathrm{t}},\ \theta_{1,1}^{\mathrm{t}}\right),\cdots,\ \boldsymbol{a}_{\mathrm{t}}\left(\phi_{N_{\mathrm{cl}},N_{\mathrm{ray}}}^{\mathrm{t}},\ \theta_{N_{\mathrm{cl}},N_{\mathrm{ray}}}^{\mathrm{t}}\right)\right]$ 分别被定义为接收阵列响应矩阵和发送阵列响应矩阵。

注意到阵列响应矢量 $\boldsymbol{a}_{\mathrm{r}}\left(\phi_{il}^{\mathrm{r}},\theta_{il}^{\mathrm{r}}\right)$ 和 $\boldsymbol{a}_{\mathrm{t}}\left(\phi_{il}^{\mathrm{t}},\theta_{il}^{\mathrm{t}}\right)$ 仅取决于天线阵列的结构。两个常用的天线阵列结构是均匀线性阵列（uniform linear array，ULA）和均匀平面阵列（uniform planar array，UPA）。尽

管本文随后给出的算法和推导结果可以适用于任意天线阵列，为了便于后续说明及性能仿真，以处在 y-z 平面上的 UPA 天线结构为例给出阵列响应矢量。若 UPA 天线阵列在 y 轴和 z 轴分别具有 W 和 H 个天线阵子（antenna element），则阵列响应矢量可以写成：

$$\boldsymbol{a}_{\mathrm{UPA}}(\phi,\theta)=\frac{1}{N}\left[1,\cdots,\mathrm{e}^{\mathrm{j}kd(m\sin\phi\sin\theta+n\cos\theta)},\cdots,\mathrm{e}^{\mathrm{j}kd((W-1)\sin\phi\sin\theta+(H-1)\cos\theta)}\right] \tag{5}$$

其中，$k=\dfrac{2\pi}{\lambda}$，d 是相邻天线阵子间的距离；$0\leqslant m\leqslant W$ 和 $0\leqslant n\leqslant H$ 分别是 y 轴和 z 轴天线阵子的索引值，且总的天线阵子数即阵列大小 $N=WH$。

3　基于有限反馈码本的模数混合预编码设计方法

对于 FDD 系统，信道状态信息（channel state information，CSI）通常只能依靠接收端的反馈获得。考虑到反馈信道一般是容量受限的，并且由于 RF 硬件的限制使得模拟移相器的相移往往是离散的[16]，因此预编码矩阵只能从有限的集合中选取。此外，由于毫米波系统的模拟/数字混合编码属性，使得码本可以分为模拟预编码码本和数字预编码码本，即双码本结构，从而降低了搜索空间，减少了复杂度。本文基于双码本结构，提出了针对单用户毫米波 MIMO 系统的有限反馈模数混合预编码及模拟预编码码本的设计方法。该方法首先要求接收端根据 CSI 及所提出的优化目标函数从模拟预编码码本中选择模拟预编码矩阵，其中模拟预编码码本的构造采用新的设计方法与准则，即基于 DFT 矩阵及其旋转矩阵设计码本；进而以 CSI 和模拟预编码矩阵从随机矢量量化（random vector quantization，RVQ）码本中选择数字预编码矩阵；最后将所选择的模拟预编码矩阵和数字预编码矩阵的索引值通过反馈信道发送回发送端。为了便于说明本文所提算法，下文首先简单介绍最佳的全数字预编码算法和参考文献[6]提出的模数混合预编码算法。

3.1　预编码算法理论设计

假设发送端已知信道状态信息，为了获得最佳预编码矩阵，对信道 $\boldsymbol{H}$ 进行降序奇异值分解（SVD），使得 $\boldsymbol{H}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{H}}$。其中，$\boldsymbol{U}\in\mathbb{C}^{N_{\mathrm{r}}\times\mathrm{rank}(\boldsymbol{H})}$ 和 $\boldsymbol{V}\in\mathbb{C}^{N_{\mathrm{t}}\times\mathrm{rank}(\boldsymbol{H})}$ 均为酉阵，$\boldsymbol{\Sigma}\in\mathbb{C}^{\mathrm{rank}(\boldsymbol{H})\times\mathrm{rank}(\boldsymbol{H})}$ 是奇异值按降序排列的对角阵，另外 $\mathrm{rank}(\boldsymbol{H})$ 为信道 $\boldsymbol{H}$ 的秩。进一步地，可以将 $\boldsymbol{\Sigma}$ 和 $\boldsymbol{V}$ 写成如下的分块矩阵形式：

$$\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_2\end{bmatrix},\quad \boldsymbol{V}=\begin{bmatrix}\boldsymbol{V}_1 & \boldsymbol{V}_2\end{bmatrix} \tag{6}$$

其中，$\boldsymbol{\Sigma}_1\in\mathbb{C}^{N_{\mathrm{s}}\times N_{\mathrm{s}}}$ 和 $\boldsymbol{V}_1\in\mathbb{C}^{N_{\mathrm{t}}\times N_{\mathrm{s}}}$。因此，对应于信道 $\boldsymbol{H}$，发射机的最佳无约束单位预编码矩阵为 $\boldsymbol{F}_{\mathrm{opt}}=\boldsymbol{V}_1$，即全数字预编码矩阵。

考虑到毫米波系统中全数字预编码难以实现，因此毫米波通信往往采用模数混合预编码。假设接收端采用基于最大似然的全数字译码算法，对于模拟/数字预编码矩阵 $(\boldsymbol{F}_{\mathrm{RF}},\boldsymbol{F}_{\mathrm{BB}})$ 的联合设计问题，参考文献[6]利用毫米波信道的空间稀疏特性，将预编码问题规划为稀疏重构问题，提出了基于正交匹配追踪（orthogonal matching pursuit，OMP）的空间稀疏预编码算法。

为了便于参考，参考文献[6]中 OMP 算法的伪代码如下。

函数：Sparse_TX_RX$(\boldsymbol{A}_{\mathrm{t}},\boldsymbol{F}_{\mathrm{opt}},N_{\mathrm{t}}^{\mathrm{RF}},\rho)$；

输入：$\boldsymbol{A}_{\mathrm{t}}$，$\boldsymbol{F}_{\mathrm{opt}}$，$N_{\mathrm{t}}^{\mathrm{RF}}$ 和 ρ；

输出：$\boldsymbol{F}_{\mathrm{RF}}$，$\boldsymbol{F}_{\mathrm{BB}}$；

$\boldsymbol{F}_{\mathrm{res}}=\boldsymbol{F}_{\mathrm{opt}}$；

$\boldsymbol{F}_{\mathrm{RF}}$=Empty；

for $r\leqslant N_{\mathrm{t}}^{\mathrm{RF}}$ do

$\boldsymbol{\Phi}=\boldsymbol{A}_{\mathrm{t}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{res}}$；

$k=\arg\min_l\left[\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}}\right]_{l,l}$；

$\boldsymbol{F}_{\mathrm{RF}}=\left[\boldsymbol{F}_{\mathrm{RF}}\middle|\boldsymbol{A}_{\mathrm{t}}^{(\mathrm{k})}\right]$；

$$\boldsymbol{F}_{\mathrm{BB}}=\left(\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\right)^{-1}\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{opt}};$$

$$\boldsymbol{F}_{\mathrm{res}}=\frac{\boldsymbol{F}_{\mathrm{opt}}-\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}}{\left\|\boldsymbol{F}_{\mathrm{opt}}-\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}};$$

end for ；

$\boldsymbol{F}_{\mathrm{BB}}=\sqrt{\rho}\dfrac{\boldsymbol{F}_{\mathrm{BB}}}{\left\|\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}}$ 归一化处理；

该算法提出了解决稀疏约束矩阵重构问题的方法。优化目标问题可以写成下列形式：

$$\min_{\tilde{\boldsymbol{F}}_{\mathrm{BB}}}\left\|\boldsymbol{F}_{\mathrm{opt}}-\boldsymbol{A}_{\mathrm{t}}\tilde{\boldsymbol{F}}_{\mathrm{BB}}\right\|_{\mathrm{F}}$$

$$\text{s.t.}\ \left\|\mathrm{diag}\left(\tilde{\boldsymbol{F}}_{\mathrm{BB}}\tilde{\boldsymbol{F}}_{\mathrm{BB}}^{*}\right)\right\|_{0}=N_{\mathrm{t}}^{\mathrm{RF}}$$

$$\left\|\boldsymbol{A}_{\mathrm{t}}\tilde{\boldsymbol{F}}_{\mathrm{BB}}\right\|_{\mathrm{F}}^{2}=N_{\mathrm{s}} \tag{7}$$

该算法首先需要给定 $\boldsymbol{F}_{\mathrm{opt}}$ 和 $\boldsymbol{A}_{\mathrm{t}}$，然后通过迭代获得模拟预编码矩阵 $\boldsymbol{F}_{\mathrm{RF}}$ 和数字预编码矩阵 $\boldsymbol{F}_{\mathrm{BB}}$。注意到，$\boldsymbol{F}_{\mathrm{BB}}=\left(\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\right)^{-1}\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{opt}}$ 实际上是基于 $\left\|\boldsymbol{F}_{\mathrm{opt}}-\boldsymbol{A}_{\mathrm{t}}\tilde{\boldsymbol{F}}_{\mathrm{BB}}\right\|_{\mathrm{F}}$ 的无约束最小二乘算法给出。从该 OMP 算法可以发现，为了获得模拟与数字预编码矩阵，需要首先给定发送阵列响应矩阵 $\boldsymbol{A}_{\mathrm{t}}$。然而，参考文献[15]已指出这需要额外的开销，将导致系统的频谱效率降低。此外，由于每次迭代均需矩阵求逆运算，复杂度仍然较高。

3.2 基于有限反馈码本的模数混合预编码设计方法步骤

对于 FDD 系统而言，其信道状态信息需要利用有限容量的反馈信道进行反馈。因此，一种可行的方法是接收端利用所获得的信道矩阵计算获得全数字预编码矩阵 $\boldsymbol{F}_{\mathrm{opt}}$，并以此重构混合模拟和基带预编码矩阵，使得 $\boldsymbol{F}_{\mathrm{opt}}\approx\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}$，且 $\boldsymbol{F}_{\mathrm{RF}}\in\mathcal{F}$，$\boldsymbol{F}_{\mathrm{BB}}\in\mathcal{W}$，然后将码本索引通过反馈链路反馈给发送端。尽管参考文献[6]给出了相关的基于反馈的预编码方案，但因其基于 OPM 算法，仍然具有开销较大、复杂度较高的问题。因此，本文提出一种新的基于反馈的预编码方法，能够显著降低系统开销及复杂度。

根据参考文献[6]的描述，一般的混合预编码设计的优化目标问题可以写成：

$$\min_{\boldsymbol{F}_{\mathrm{RF}},\boldsymbol{F}_{\mathrm{BB}}}\left\|\boldsymbol{F}_{\mathrm{opt}}-\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}$$

$$\text{s.t.}\ \ \boldsymbol{F}_{\mathrm{RF}}\in\mathcal{F}$$

$$\left\|\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}^{2}=N_{\mathrm{s}} \tag{8}$$

又根据参考文献[15]，可以首先考虑固定模拟预编码器 $\boldsymbol{F}_{\mathrm{RF}}$，设计数字预编码器 $\boldsymbol{F}_{\mathrm{BB}}$。因此，基于无约束的最小二乘法，可以得到 $\boldsymbol{F}_{\mathrm{BB}}=\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{opt}}$。借助该等式，式（8）所描述的优化问题可以进一步写成：

$$\min_{\boldsymbol{F}_{\mathrm{RF}}}\left\|\boldsymbol{F}_{\mathrm{opt}}-\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{opt}}\right\|_{\mathrm{F}}$$

$$\text{s.t.}\ \ \boldsymbol{F}_{\mathrm{RF}}\in\mathcal{F}$$

$$\left\|\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{opt}}\right\|_{\mathrm{F}}^{2}=N_{\mathrm{s}} \tag{9}$$

从式（9）可以看出，只须单独对模拟预编码矩阵进行设计即可，无需在模拟预编码矩阵和数字预编码矩阵之间进行迭代优化。尽管性能可能会有所降低，但显著地降低了复杂度，这对于能量受限的移动终端来说尤其具有意义。

为了找到最优的模拟预编码矩阵 $\boldsymbol{F}_{\mathrm{RF}}$ 和数字预编码矩阵 $\boldsymbol{F}_{\mathrm{BB}}$，所采取的方法是将模拟和数字预编码矩阵的搜索分别限制在两组预定义的码本 F 和 $\mathcal{W}$ 内。注意到所采用的模拟预编码码本 F 与参考文献[6]所提的基于 $\boldsymbol{A}_{\mathrm{t}}$ 的码本不同，能够显著降低开销。这是因为 $\boldsymbol{A}_{\mathrm{t}}$ 是随信道 $\boldsymbol{H}$ 不断变化的，这需要不停更新码本；而所提出的码本 F 设计方法无需 $\boldsymbol{A}_{\mathrm{t}}$ 信息，可以在很长时间内保持不变。在已知最佳预编码矩阵 $\boldsymbol{F}_{\mathrm{opt}}$ 的前提下，根据式（9）从模拟预编码码本 F 里选择近似最优的模拟预编码矩阵 $\boldsymbol{F}_{\mathrm{RF}}$；再基于最小二乘法获得最优数字预编码矩阵 $\boldsymbol{F}_{\mathrm{BB}}^{\mathrm{opt}}$；然后基于最小弦距离准则，从数字预编码码本 $\mathcal{W}$ 中选择与最佳基带预编码器 $\boldsymbol{F}_{\mathrm{BB}}^{\mathrm{opt}}$ 最接近的矩阵，从而获得数字预编码矩阵 $\boldsymbol{F}_{\mathrm{BB}}$。所提算法的伪码如下所示。

输入：$\mathcal{F}$，$\boldsymbol{F}_{\mathrm{opt}}$，$\mathcal{W}$，$N_{\mathrm{t}}^{\mathrm{RF}}$ 和 ρ；

输出：$\boldsymbol{F}_{\mathrm{RF}}$，$\boldsymbol{F}_{\mathrm{BB}}$；

$\boldsymbol{F}_{\mathrm{RF}}$=Empty；

$\boldsymbol{F}_{\mathrm{BB}}$=Empty；

$\boldsymbol{F}_{\mathrm{RF}}=\arg\min_{\tilde{\boldsymbol{F}}_{\mathrm{RF}}\in\mathcal{F}}\left\|\boldsymbol{F}_{\mathrm{opt}}-\tilde{\boldsymbol{F}}_{\mathrm{RF}}\tilde{\boldsymbol{F}}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{opt}}\right\|$；

$\boldsymbol{F}_{\mathrm{BB}}^{\mathrm{opt}}=\left(\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\right)^{-1}\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{opt}}$；

$\boldsymbol{F}_{\mathrm{BB}}=\arg\max_{\tilde{\boldsymbol{F}}_{\mathrm{BB}}\in\mathcal{W}}\left\|\tilde{\boldsymbol{F}}_{\mathrm{BB}}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{BB}}^{\mathrm{opt}}\right\|$；

$\boldsymbol{F}_{\mathrm{BB}}=\sqrt{\rho}\dfrac{\boldsymbol{F}_{\mathrm{BB}}}{\left\|\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}}$ 归一化处理；

该算法的主要思想是将模拟/数字预编码矩阵的形式分别限定在基于 DFT 矩阵的模拟预定义矩阵 F 和 RVQ 码本 $\mathcal{W}$ 里，再通过上述优化问题对模拟预编码矩阵和数字预编码矩阵分别进行码字搜索，使用该算法可以显著降低系统复杂度。

3.3　模拟预定义码本的构造

根据第 3.2 节所示，该算法的关键在于提出一种新的模拟预编码码本构造方法。相比于参考文献[8]的 RVQ 码本构造方法，此码本可以获得更好的性能。所提码本的构造方法如下。

首先设 L 为码本 F 的大小，且要求 $L=2^b$，其中，b 为反馈比特数；码本 F 可表示为 $\mathcal{F}=\{\boldsymbol{\Phi}_1,\boldsymbol{\Phi}_2,\cdots,\boldsymbol{\Phi}_L\}$。

模拟预编码码本的设计准则：由码字 $\boldsymbol{\Phi}_i$ 的列向量所生成的空间记<$\boldsymbol{\Phi}_i$>，它是一个 N_{t} 维空间中的 $N_{\mathrm{t}}^{\mathrm{RF}}$ 维子空间。其中，N_{t} 和 $N_{\mathrm{t}}^{\mathrm{RF}}$ 分别是发射端的天线数和射频链路数。这样，<$\boldsymbol{\Phi}_i$>与<$\boldsymbol{\Phi}_j$>之间存在 $N_{\mathrm{t}}^{\mathrm{RF}}$ 个角度，记这 $N_{\mathrm{t}}^{\mathrm{RF}}$ 个角度分别为 $\omega_{ij,1}\leqslant\omega_{ij,2}\leqslant\cdots\leqslant\omega_{ij,N_{\mathrm{t}}^{\mathrm{RF}}}$；同时，记 $\lambda_{ij,1}\leqslant\lambda_{ij,2}\leqslant\cdots\leqslant\lambda_{ij,N_{\mathrm{t}}^{\mathrm{RF}}}$ 为矩阵 $\boldsymbol{\Phi}_i^*\boldsymbol{\Phi}_j\boldsymbol{\Phi}_j^*\boldsymbol{\Phi}_i$ 的特征值，则它们的关系为 $\omega_{ij,\mathrm{r}}=\arccos\left(\sqrt{\lambda_{ij,\mathrm{r}}}\right)$（$r=1,\cdots,N_{\mathrm{t}}^{\mathrm{RF}}$）。当索引 $r>N_{\mathrm{t}}/2$ 时，有 $\omega_{ij,\mathrm{r}}=0$，即当索引数超过天线数的一半时，两个子空间之间的角度会出现重合；由此，当 $N_{\mathrm{t}}^{\mathrm{RF}}\leqslant N_{\mathrm{t}}/2$ 时，选择码字使得 $\min_{1\leqslant i<j\leqslant L}\omega_{ij,\mathrm{r}}$ 越大越好；而当 $N_{\mathrm{t}}^{\mathrm{RF}}>N_{\mathrm{t}}/2$ 时，选择码字使得 $\min_{1\leqslant i<j\leqslant L}\omega_{ij,\mathrm{t}}$，其中 $t=N_{\mathrm{t}}/2$。

该设计准则实际上可以最大化任意一对码字的列向量所生成的子空间之间的最小角度，可以降低码本的失真度，即提高码本质量。根据仿真表明[17]，使用该设计准则可以更好地逼近真实的波束成形矢量。

模拟预编码码本 F 的构造方法如下：令 $L=L_1\times L_2\times\cdots\times L_{\mathrm{s}}$，其中，$L_1,L_2,\cdots,L_{\mathrm{s}}$ 为待定的正整数；天线数为 N_{t}，则码本 F 采用如下的设计方法：$\mathcal{F}=\{\boldsymbol{\Phi}_1,\boldsymbol{\Phi}_2,\cdots,\boldsymbol{\Phi}_L\}$；其中 $\boldsymbol{\Phi}_l=\theta_1^{l_1-1}\theta_2^{l_2-1}\cdots\ \theta_{\mathrm{s}}^{l_{\mathrm{s}}-1}\boldsymbol{\Phi}_0$，$1\leqslant l_i\leqslant L_i, i=1,\cdots,s$，而 θ_i 是构造的旋转矩阵：

$$\theta_i=\begin{pmatrix}\mathrm{e}^{\mathrm{j}\frac{2\pi}{L_i}u_{i1}} & 0 & \cdots & 0\\ 0 & \mathrm{e}^{\mathrm{j}\frac{2\pi}{L_i}u_{i2}} & \cdots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & 0 & \cdots & \mathrm{e}^{\mathrm{j}\frac{2\pi}{L_i}u_{iN_{\mathrm{t}}}}\end{pmatrix}_{i=1,2,\cdots,s}\tag{10}$$

其中，$0\leqslant u_{i1},u_{i2},\cdots,u_{iN_{\mathrm{t}}}\leqslant L_i-1$ 为 N_{t} 个待定的整数，并记为 $\boldsymbol{U}_i=\begin{bmatrix}u_{i1} & u_{i2} & \cdots & u_{iN_{\mathrm{t}}}\end{bmatrix}$，则有 $\boldsymbol{U}=\begin{bmatrix}\boldsymbol{U}_1^{\mathrm{T}} & \boldsymbol{U}_2^{\mathrm{T}} & \cdots & \boldsymbol{U}_{\mathrm{s}}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$；$\boldsymbol{\Phi}_0$ 为 $N_{\mathrm{t}}\times N_{\mathrm{t}}^{\mathrm{RF}}$ 维的酉矩阵，即满足 $\boldsymbol{\Phi}_0^*\boldsymbol{\Phi}_0=\boldsymbol{I}_{\mathrm{r}}$（$\boldsymbol{I}_N$ 为 N 维的单位矩阵），$N_{\mathrm{t}}^{\mathrm{RF}}$ 为码字的秩，所以可以从 $N_{\mathrm{t}}\times N_{\mathrm{t}}$ 维的 DFT 矩阵中选取前 $N_{\mathrm{t}}^{\mathrm{RF}}$ 列来构造 $\boldsymbol{\Phi}_0$；而 $u_{i1},u_{i2},\cdots,u_{iN_{\mathrm{t}}}$ 的大小也是根据上述码本的设计准则即最小角度最大化的方法来确定。以最简单的情况为例，即当 $L=L_1$ 时，有 $\boldsymbol{U}=\boldsymbol{U}_1=\begin{bmatrix}u_{11} & u_{12} & \cdots & u_{1N_{\mathrm{t}}}\end{bmatrix}$，则 $\boldsymbol{U}$ 的选取方法如下所示。注意，该算法可直接推广到 $L=L_1\times L_2\times\cdots\times L_{\mathrm{s}}$ 的一般情况。则 $u_{11},u_{12},\cdots,u_{1N_{\mathrm{t}}}$ 的选取算法的伪码如下。

输入：$\boldsymbol{\Phi}_0$，θ_1；

输出：$\boldsymbol{U}$；

$\boldsymbol{\Phi}_l=\theta_1^{l_1-1}\boldsymbol{\Phi}_0$，$1\leqslant l_1\leqslant L_1$；

$\boldsymbol{\Phi}_l=\theta_1^{l_1-1}\boldsymbol{\Phi}_0$，$1\leqslant l_1\leqslant L_1$；

$\boldsymbol{U}=\arg_{0\leqslant u_{11},u_{12},\cdots,u_{1N_{\mathrm{t}}}\leqslant L_i-1}\max\{\min\arccos(\mathrm{sqrt}(\boldsymbol{\Phi}_i^{\mathrm{H}}\boldsymbol{\Phi}_j\boldsymbol{\Phi}_j^{\mathrm{H}}\boldsymbol{\Phi}_i))\}$，$1\leqslant i<j\leqslant L$；

例如当系统参数为 $L=16$，$N_{\mathrm{t}}=8$，$N_{\mathrm{t}}^{\mathrm{RF}}=4$ 时，码本可以设计为：取 $\boldsymbol{\Phi}=\left(\mathrm{e}^{\mathrm{j}\frac{2\pi}{8}kn}\right)_{8\times8}$，

$k=0,1,\cdots,7$，$n=0,1,\cdots,7$，则$\boldsymbol{\Phi}$为 8 阶的 DFT 矩阵，记φ_j为$\boldsymbol{\Phi}$的第 j 列，所以$\boldsymbol{\Phi}_0$可取$\boldsymbol{\Phi}_0=[\varphi_1 \quad \varphi_2 \quad \varphi_3 \quad \varphi_4]$。若令$L=L_1=16$时，可取$u_{11}=0$，$u_{12}=1$，$u_{13}=7$，$u_{14}=5$，$u_{15}=4$，$u_{16}=14$，$u_{17}=15$，$u_{18}=12$，即$\boldsymbol{U}=[0\ 1\ 7\ 5\ 4\ 14\ 15\ 12]$；所以码本为$\mathcal{F}=\left\{\theta_1^{l_1}\boldsymbol{\Phi}_0;0\leqslant l_1\leqslant 15\right\}$。类似地，由于有$L=16=4\times 4$，则也可令$L_1=4$，$L_2=4$，此时可以获取与上述$L=L_1=16$不同的另外一个码本。该码本首先获得$U_1$和$U_2$，则有$\boldsymbol{U}=\begin{bmatrix}0&4&3&3&4&4&3&3\\0&1&4&3&2&3&4&3\end{bmatrix}$，所以此时码本为$\mathcal{F}=\left\{\theta_1^{l_1}\theta_2^{l_2}\boldsymbol{\Phi}_0;0\leqslant l_1\leqslant 3,0\leqslant l_2\leqslant 3\right\}$。

4 仿真结果和分析

在本节中，将通过一些仿真来表明所提算法的有效性。具体的仿真参数如下：考虑一个单用户毫米波 MIMO 系统，发送端配备有8×8 UPA（$N_{\mathrm{t}}=64$）和$N_{\mathrm{t}}^{\mathrm{RF}}=4$射频链路数，接收端配备有$4\times 4$ UPA（$N_{\mathrm{r}}=16$）和$N_{\mathrm{r}}^{\mathrm{RF}}=4$射频链路数。信道一共有 8 个簇，且每个簇的路径数均为 10，即$N_{\mathrm{cl}}N_{\mathrm{ray}}=80$。假设信道每个簇的路径增益$\alpha_{il}$服从方差为$\sigma_{\alpha,i}^2=1$的高斯分布。并假设到达角和出发角的方位角在$[0,2\pi]$内均匀分布，而俯仰角在$\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$内均匀分布。设置噪声方差$\sigma^2=1$，$\mathrm{SNR}=\frac{P\sigma_\alpha^2}{N_{\mathrm{s}}}$，$N_{\mathrm{s}}=1$。

在仿真中，除了本文所提的算法外，还对比了其他 6 种预处理方法，分别是：发射机已知完全信道状态信息下，分别采用全数字预编码和基于 OMP 的模/数混合预编码方案；发射机已知反馈的信道状态信息，即发射机通过接收机反馈获得关于信道矩阵$\boldsymbol{H}$的信息，分别采用全数字预编码和基于 OMP 的模/数混合预编码方案；发射机获得接收机反馈的关于全数字预编码矩阵$\boldsymbol{F}_{\mathrm{opt}}$的信息，直接采用该反馈预编码矩阵的全数字预编码和基于 OMP 的模/数混合预编码方案。

如图 2 所示，假定用于反馈信道矩阵$\boldsymbol{H}$或者最佳数字预编码矩阵$\boldsymbol{F}_{\mathrm{opt}}$的反馈比特数为$b$=10 bit；本文提出的双码本模/数混合预编码设计方法中，设定模拟预编码码本的反馈比特数为b_1=4 bit，数字预编码码本的反馈比特数为b_2=6 bit。当反馈链路的总比特数相同，且$N_{\mathrm{s}}=1$时，对其频谱效率的性能进行比较。可以看出，所提方法的频谱效率与直接反馈信道矩阵$\boldsymbol{H}$的最佳全数字预编码方法的性能相比，性能提高了约 2.8 dB；同时亦略高于基于反馈信道矩阵$\boldsymbol{H}$的 OMP 模/数混合预编码方法和基于反馈$\boldsymbol{F}_{\mathrm{opt}}$的最佳全数字预编码方法；但是低于基于反馈最佳数字预编码矩阵$\boldsymbol{F}_{\mathrm{opt}}$的 OMP 模/数混合预编码方法。需要注意的是，在仿真中，不管是基于反馈信道矩阵$\boldsymbol{H}$还是最佳数字预编码矩阵$\boldsymbol{F}_{\mathrm{opt}}$，OMP 模/数混合预编码方法均优于全数字预编码方法，这是因为在仿真中假设发送阵列响应矩阵$\boldsymbol{A}_{\mathrm{t}}$在发射端是完全已知的。但是在实际中，$\boldsymbol{A}_{\mathrm{t}}$完全已知显然难以满足，此外发射机需要实时更新$\boldsymbol{A}_{\mathrm{t}}$也将导致系统开销的增加，由此将导致频谱效率的降低。最后，尽管都反馈 10 bit 信息，但是由于所提算法采用双码本结构，搜索空间最大只有2^6=64，而单码本结构的搜索空间为2^{10}=1 024，这能显著降低终端的复杂度。

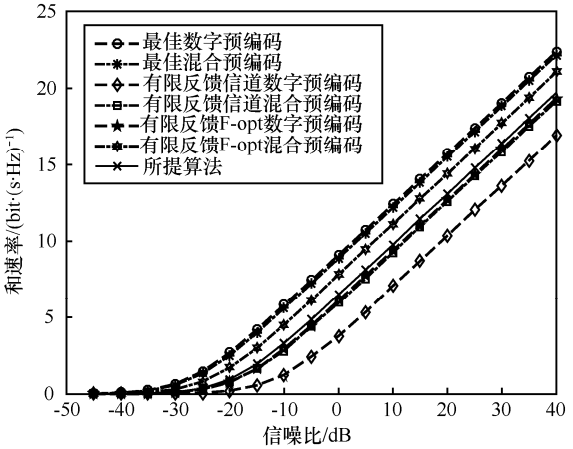


图 2 系统和速率随着信噪比的变化曲线

与图 2 的仿真略有不同的是，在图 3 的仿真中，将总的反馈比特数提升至 8 bit，其中所提算法的模拟预编码码本的反馈比特数为 b_1=4 bit，数字预编码码本的反馈比特数为 b_2=4 bit。从图 3 可以观察到与图 2 类似的结论。此外，对比图 2 和图 3 可以看到，尽管图 3 的反馈比特数降至 8 bit，但所达到的频谱效率与图 2 基本接近，因此在所给的系统参数下，实际上 8 bit 已足够满足系统需求。

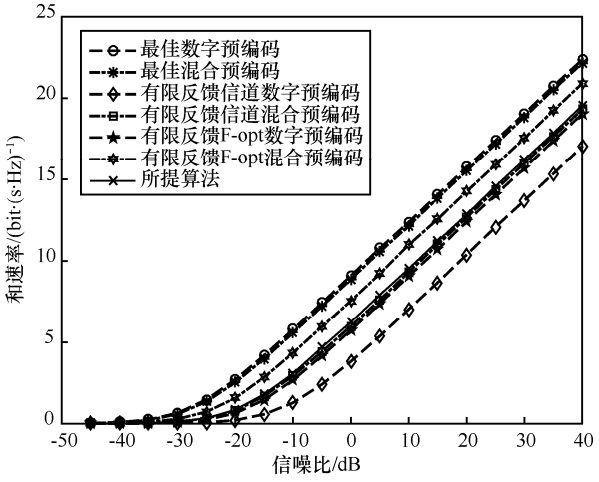


图 3　系统和速率随着信噪比的变化曲线

在图 4 和图 5 中，将并行传输的数据流增加到 2，即 $N_s=2$，对比各种预处理方法的频谱效率性能。与 $N_s=1$ 不同的是，所提算法的性能优于基于反馈信道矩阵 $\boldsymbol{H}$ 的全数字预编码方法和基于反馈 $\boldsymbol{F}_{\text{opt}}$ 的全数字预编码方法，但是略低于基于反馈信道矩阵 $\boldsymbol{H}$ 的 OMP 模/数混合预编码方法和基于反馈 $\boldsymbol{F}_{\text{opt}}$ 的 OMP 模/数混合预编码方法。但是考虑到实际系统中 $\boldsymbol{A}_t$ 在发射端难以完全已知以及接收端码字搜索的复杂性等情形，所提算法仍具有较强的竞争力。

5　结束语

针对单用户毫米波 MIMO 系统的下行链路通信，提出了基于有限反馈码本的低复杂度混合模拟/数字预编码方法。由于该预编码方法采用了基

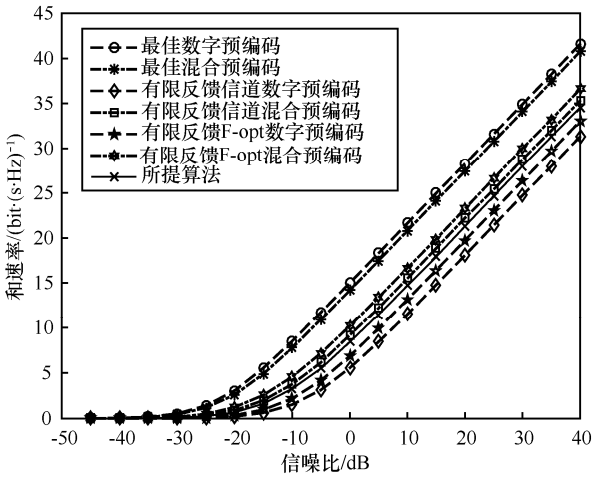


图 4　系统和速率随着信噪比的变化曲线

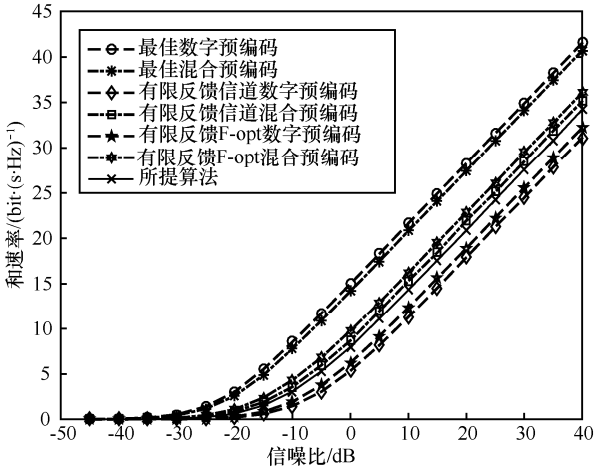


图 5　系统和速率随着信噪比的变化曲线

于 DFT 矩阵和旋转矩阵的新模拟码本构造方法，保证了系统性能。同时又将模数预编码矩阵的混合优化问题转变为模拟预编码矩阵的单独优化问题，从而无须迭代优化，即可使得码字搜索在模拟域和数字域中分别进行，降低了复杂度。通过仿真可以观察到，本文所提出的算法能够在性能和复杂度上获得较好的均衡。未来可将此算法扩展到多用户多小区的场景下，并针对该场景对算法进行进一步的优化研究。

参考文献：

[1] ANDREWS J G, BUZZI S, WAN C, et al. What will 5G be?[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(6): 1065-1082.

[2] PI Z, KHAN F. An introduction to millimeter-wave mobile

broadband systems[J]. IEEE Communications Magazine, 2011, 49(6): 101-107.
[3] RAPPAPORT T S, SUN S, MAYZUS R, et al. Millimeter wave mobile communications for 5G cellular: it will work![J]. IEEE Access, 2013, 1(1): 335-349.
[4] RAPPAPORT T S, JR R W H, DANIELS R C, et al. Millimeter wave wireless communications[M]. Englewood: Prentice Hall, 2015.
[5] BOCCARDI F, HEATH R W, LOZANO A, et al. Five disruptive technology directions for 5G[J]. IEEE Communications Magazine, 2014, 52(2): 74-80.
[6] AYACH O E, RAJAGOPAL S, ABU-SURRA S, et al. Spatially sparse precoding in millimeter wave MIMO systems[J]. IEEE Transactions on Wireless Communications, 2014, 13(3): 1499-1513.
[7] ALKHATEEB A, HEATH R W, LEUS G. Achievable rates of multi-user millimeter wave systems with hybrid precoding[C]//IEEE International Conference on Communication Workshop, June 8-12, 2015, Torino, Italy. Piscataway: IEEE Press, 2015: 1232-1237.
[8] ALKHATEEB A, LEUS G, HEATH R W. Limited feedback hybrid precoding for multi-user millimeter wave systems[J]. IEEE Transactions on Wireless Communications, 2014, 14(11): 6481-6494.
[9] MÉNDEZ-RIAL R, RUSU C, GONZÁLEZ-PRELCIC N, et al. Hybrid MIMO architectures for millimeter wave communications: phase shifters or switches?[J]. IEEE Access, 2015(4): 247-267.
[10] HEATH R W, GONZÁLEZ-PRELCIC N, RANGAN S, et al. An overview of signal processing techniques for millimeter wave MIMO systems[J]. IEEE Journal of Selected Topics in Signal Processing, 2015, 10(3): 436-453.
[11] ALKHATEEB A, MO J, GONZALEZ-PRELCIC N, et al. MIMO precoding and combining solutions for millimeter-wave systems[J]. IEEE Communications Magazine, 2014, 52(12): 122-131.
[12] ALKHATEEB A, AYACH O E, LEUS G, et al. Channel estimation and hybrid precoding for millimeter wave cellular systems[J]. IEEE Journal of Selected Topics in Signal Processing, 2014, 8(5): 831-846.
[13] HAN S, CHIH-LIN I, XU Z, et al. Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G[J]. IEEE Communications Magazine, 2015, 53(1): 186-194.
[14] BRADY J, BEHDAD N, SAYEED A M. Beamspace MIMO for millimeter-wave communications: system architecture, modeling, analysis, and measurements[J]. IEEE Transactions on Antennas & Propagation, 2013, 61(7): 3814-3827.
[15] YU X, SHEN J C, ZHANG J, et al. Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems[J]. IEEE Journal of Selected Topics in Signal Processing, 2016, 10(3): 485-500.
[16] VENKATESWARAN V, VEEN A J V D. Analog beamforming in mimo communications with phase shift networks and online channel estimation[J]. IEEE Transactions on Signal Processing, 2010, 58(8): 4131-4143.
[17] WANG H Q, ZHAO Z J. A MIMO system with finite-bit feedback based on fixed constellations[J]. Science China, 2013, 56(6): 1-14.
[18] DAI L, GAO X, QUAN J, et al. Near-optimal hybrid analog and digital precoding for downlink mmWave massive MIMO systems[R]. 2015.
[19] LI A, MASOUROS C, LI A, et al. Hybrid precoding and combining design for millimeter-wave multi-user MIMO based on SVD[C]// ICC 2017-2017 IEEE International Conference on Communications, May 21-25, 2017, Paris, France. Piscataway: IEEE Press, 2017: 1-6.
[20] LI C, LIU P, ZOU C, et al. Spectral-efficient cellular communications with coexistent one- and two-hop transmissions[J]. IEEE Transactions on Vehicular Technology, 2016, 65(8): 6765-6772.
[21] LI C, ZHANG S, LIU P, et al. Overhearing protocol design exploiting intercell interference in cooperative green networks[J]. IEEE Transactions on Vehicular Technology, 2016, 65(1): 441-446.
[22] LI C, YANG H J, SUN F, et al. Multiuser overhearing for cooperative two-way multiantenna relays[J]. IEEE Transactions on Vehicular Technology, 2016, 65(5): 3796-3802.
[23] LI C, SUN F, CIOFFI J M, et al. Energy efficient MIMO relay transmissions via joint power allocations[J]. IEEE Transactions on Circuits & Systems II Express Briefs, 2014, 61(7): 531-535.
[24] LI C, YANG H J, SUN F, et al. Adaptive overhearing in two-way multi-antenna relay channels[J]. IEEE Signal Processing Letters, 2015, 23(1): 117-120.

[作者简介]

尤若楠（1992−），女，杭州电子科技大学通信工程学院硕士生，主要研究方向为毫米波通信。

潘鹏（1983−），男，杭州电子科技大学通信工程学院副教授，主要研究方向为 MIMO 及大规模 MIMO 预编码和容量分析、毫米波通信。

张丹（1993−），女，杭州电子科技大学通信工程学院硕士生，主要研究方向为毫米波通信。

王海泉（1964−），男，杭州电子科技大学通信工程学院教授、博士生导师，主要研究方向为无线通信、多天线系统、信号检测、信息论等。

研究与开发

基于信任度的可变门限能量检测算法

肖洁，陈跃斌，陈楚天，郑婷，钱继武
（云南民族大学电气信息工程学院，云南 昆明 650500）

摘　要：针对随机的概率式频谱感知数据篡改（spectrum sensing data falsification，SSDF）攻击，提出基于信任度的可变门限能量检测算法。首先比较实际融合值与融合中心上、下边界值的关系，更新可变门限，且通过系统给定的虚警概率和漏检概率确定上、下边界值；其次采用基于正确感知次数和总感知次数比值确定信任值的软融合方法。仿真结果表明，与传统固定门限相比，算法抵御攻击的同时不仅能够降低虚警概率和漏检概率，同时可以提高系统检测概率。

关键词：认知无线电；能量检测；概率式 SSDF 攻击；可变门限
中图分类号：TN925
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018203

Variable threshold energy detection algorithm based on trust degree

XIAO Jie, CHEN Yuebin, CHEN Chutian, ZHENG Ting, QIAN Jiwu
School of Electrical and Information Technology, Yunnan Minzu University, Kunming 650500, China

Abstract: Aiming at stochastic probabilistic SSDF attacks, a variable threshold energy detection algorithm based on trust degree was proposed. Firstly, the variable threshold was updated by comparing the actual fusion value with the upper and lower boundary values of the fusion center. The upper and lower boundary values were determined through the given false alarm probability and missed probability. Secondly, a soft fusion method was used based on the ratio of correct perception times and the total number of times to update the trust value. Simulation results show that, compared with the traditional fixed threshold, the proposed algorithm can not only reduce the false alarm and missed detection probability, but also improve the detection probability of the system.

Key words: cognitive radio, energy detection, probabilistic spectrum sensing data falsification attack, variable threshold

1　引言

为了解决频谱资源紧缺问题，认知无线电（cognitive radio，CR）技术被提出，受到国内外研究学者的高度关注。CR 技术在不影响主用户（primary user，PU）正常工作的情况下动态感知

收稿日期：2017−12−22；修回日期：2018−06−08
通信作者：陈跃斌，cybuestc@sina.com
基金项目：国家自然科学基金资助项目（No.61261022）；云南民族大学创新团队项目
Foundation Items: The National Natural Science Foundation of China (No.61261022), Innovation Team Project of Yunnan Minzu University

空闲频谱[1]。作为 CR 技术的首要环节——频谱感知，感知结果的好坏直接影响整个认知无线电网络的性能。协作频谱感知通过融合多个认知用户（cognitive user，CU）的感知数据，提高频谱感知的精确性，消除单用户感知结果的不确定性和误差[2]。但协作感知过程存在着诸多风险[3-4]，SSDF（spectrum sensing data falsification，频谱感知数据篡改）攻击是其中之一，即恶意用户（malicious user, MU）通过给融合中心（fusion center，FC）发送虚假的数据以致其做出错误的判决结果[5-6]。

目前，抵御 SSDF 攻击的协作频谱感知方案较多。参考文献[7]分析 SSDF 攻击下协作频谱感知的性能限制问题并提出最优的攻击者跟踪策略，但未考虑最优策略的计算复杂度。参考文献[8]提出通过加权系数消除来自恶意用户的虚假信息，从而实现抑制恶意用户的协作检测技术，但此方法未考虑概率式 SSDF 攻击。参考文献[9]提出攻击意识的协作频谱感知方案抵御 SSDF 攻击，该方案对攻击强度进行评估，用硬判决准则得到使贝叶斯风险最小的 K 值，虽然分析了攻击强度，但未考虑概率式 SSDF 攻击。参考文献[10]提出基于改进的软融合能量检测算法抵御 SSDF 攻击方案，通过建立用户的历史服务质量信誉机制，并利用不同用户的平均信誉度合理分配权重，减少恶意用户对融合结果的影响，同样此方案也未考虑概率式 SSDF 攻击。参考文献[11]在参考文献[10]基础上针对概率式 SSDF 攻击采用协方差检测方法，基于改进的软融合方法中提出衰减因子，改善了低信噪比下的检测性能。参考文献[12]提出一种基于贝叶斯模型的协作频谱感知方案抵御 SSDF 攻击，通过比较融合值与固定门限的大小得到判决结果。虽然是抵御概率式 SSDF 攻击，但不能有效抵御攻击，原因是当恶意用户发动攻击时，融合值不断变化，系统需要门限为适应融合值的变化而变化，确保得到精确的判决结果，以有效抵御概率式 SSDF 攻击。

为了解决参考文献[12]中的问题，本文提出了一种基于信任度的可变门限策略（variable threshold energy detection based on reputation，RVTED）抵御概率式 SSDF 攻击。通过分析认知用户与以往提供感知数据的行为表现，将信任值看成正确感知次数与总感知次数的比值，给每个认知用户分配相应权值。首先攻击者发送概率式 SSDF 攻击，根据融合中心融合值的变化改变判决门限确保系统得到更精确的判决；其次在无攻击者的情况下，通过给定的较小的虚警概率和漏检概率分别得到融合值的上边界值和下边界值；最后通过比较真实融合值与上下边界值的大小关系进而确定系统门限大小，得到最终的判决结果。

2 系统模型和攻击问题

2.1 系统模型

存在多个恶意用户的协作频谱感知模型如图 1 所示，认知无线电网络模型中存在一个 FC、一个 PU 和 M 个 CU。融合中心融合每个认知用户的本地感知结果并做出最终判决，判断 PU 是否工作。第 i 个认知用户的二元假设检验：

$$y_i(n)=\begin{cases}\eta_i(n), H_0\\ h_i s(n)+\eta_i(n), H_1\end{cases} \quad i=1,2,\cdots,M \qquad (1)$$

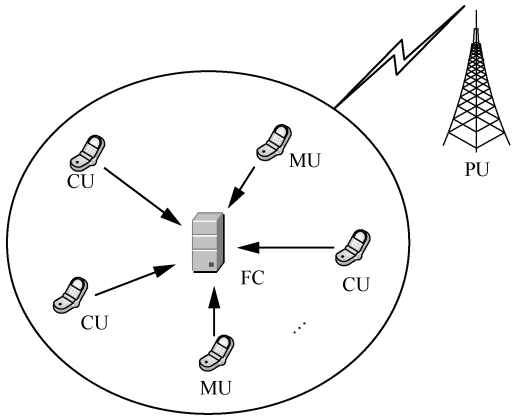


图 1 存在多个恶意用户的协作频谱感知模型

H_1 代表假设 PU 信号存在并占用频段，H_0 代表假设 PU 不存在频段空闲情况，n 表示采样的样

本数量，$y_i(n)$是第i个 CU 接收到的信号，h_i是 PU 与第i个 CU 之间进行信号传输的信道增益，$s(n)$表示 PU 信号，$\eta_i(n)$是第i个 CU 服从均值为零、方差为σ_i^2的加性高斯白噪声。$s(n)$和$\eta_i(n)$是实信号且相互独立。

2.2　能量检测

能量检测是认知无线电技术最常用的频谱感知方法。因其不需知道 PU 的先验信息，只需计算所有采样点的能量，与预先设定的门限进行比较，得到检测结果，是目前最有效的检测未知信号的方法。

假设每个 CU 的本地能量检测采样样本为N，则第i个 CU 的本地能量检测统计量Y_i为：

$$Y_i = \frac{1}{N}\sum_{n=1}^{N}\left|y_i(n)\right|^2 \tag{2}$$

由中心极限定理，当N足够大（$N \geqslant 50$）时，Y_i近似服从高斯分布[11]：

$$\begin{cases} H_0 : Y_i \sim N\left(\sigma_i^2, 2\sigma_i^4/N\right) \\ H_1 : Y_i \sim N\left((1+\gamma_i)\sigma_i^2, 2(2\gamma_i+1)\sigma_i^4/N\right) \end{cases} \tag{3}$$

其中，γ_i代表第i个认知用户接收信噪比。

根据式（3），第i个 CU 的本地检测概率$p_{\mathrm{d}}(i)$、本地虚警概率$p_{\mathrm{f}}(i)$为：

$$p_{\mathrm{d}}(i) = \Pr(Y_i \geqslant \lambda_i / H_1) = Q\left(\left(\frac{\lambda_i}{\sigma_i^2} - \gamma_i - 1\right)\sqrt{\frac{N}{4\gamma_i+2}}\right) \tag{4}$$

$$p_{\mathrm{f}}(i) = \Pr(Y_i \geqslant \lambda_i / H_0) = Q\left(\left(\frac{\lambda_i}{\sigma_i^2} - 1\right)\sqrt{\frac{N}{2}}\right) \tag{5}$$

其中，λ_i是第i个 CU 的本地判决门限，$Q(x)$为互补误差函数，且$Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^{\infty} \mathrm{e}^{-t^2/2}\mathrm{d}t$。

采用纽曼皮尔逊准则，推导式（5）得到本地判决门限λ_i：

$$\lambda_i = \left(Q^{-1}\left(p_{\mathrm{f}}(i)\right)\cdot\sqrt{N/2} + 1\right)\sigma_i^2 \tag{6}$$

2.3　概率式 SSDF 攻击过程

假设恶意用户的攻击强度Δ（偏离本地感知能量的值）、攻击概率p_a，第i个 CU 本地判决结果为d_i。协作频谱感知过程，每个 MU 以概率p_a发送 SSDF 攻击导致融合中心做出错误的判决结果。具体情况如下[9]：

（1）当 MU 不发送攻击，MU 将发送正确的能量值Y_i和本地判决结果d_i到融合中心；

（2）当存在 PU 信号占用频段的情况下（$Y_i > \lambda_i$），MU 以概率p_a发送“0”攻击，MU 将发送虚假的能量值$Y_i - \Delta$代替Y_i和本地判决结果d_i到融合中心；

（3）当不存在 PU 信号频段空闲的情况下（$Y_i < \lambda_i$），MU 以概率p_a发送“1”攻击，MU 将发送虚假的能量值$Y_i + \Delta$代替Y_i和本地判决结果d_i到融合中心。

因此第i个 CU 发送到融合中心的真实本地感知数据Z_i为：

$$Z_i = \begin{cases} Y_i, & \text{无攻击} \\ Y_i - \Delta, & \text{“0”攻击} \\ Y_i + \Delta, & \text{“1”攻击} \end{cases} \tag{7}$$

融合中心采用线性加权融合认知用户的本地感知数据，融合中心值为：

$$Z_{\mathrm{c}} = \sum_{i=1}^{M} w_i \cdot Z_i \tag{8}$$

其中，w_i是第i个 CU 的本地感知能量的加权系数且$\sum_{i=1}^{M} w_i = 1$。

3　基于可变门限策略抵御 SSDF 攻击

3.1　MU 不发送攻击

Z_{c}是 MU 不发动攻击时融合中心的值，由于Z_i服从高斯分布，则根据式（7）、式（8），Z_{c}也服从高斯分布：

$$\begin{cases} H_0 : Z_{\mathrm{c}} \sim N\left(\sum_{i=1}^{M} w_i\sigma_i^2, 2\sum_{i=1}^{M} w_i^2\sigma_i^4/N\right) \\ H_1 : Z_{\mathrm{c}} \sim N\left(\sum_{i=1}^{M} w_i(1+\gamma_i)\sigma_i^2, 2\sum_{i=1}^{M} w_i^2(1+2\gamma_i)\sigma_i^4/N\right) \end{cases} \tag{9}$$

由式（7）~式（9），MU 不发动攻击时的全局虚警概率p_{f}和全局漏检概率p_{m}为：

$$p_{\mathrm{f}}=\Pr(Z_{\mathrm{c}}\geqslant\lambda_{\mathrm{c}}/H_0)=Q\left(\frac{\lambda_{\mathrm{c}}-\sum_{i=1}^{M}w_i\sigma_i^2}{\sqrt{2\sum_{i=1}^{M}w_i^2\sigma_i^4\Big/N}}\right)\tag{10}$$

$$p_{\mathrm{m}}=\Pr(Z_{\mathrm{c}}>\lambda_{\mathrm{c}}/H_1)=1-Q\left(\frac{\lambda_{\mathrm{c}}-\sum_{i=1}^{M}w_i(1+\gamma_i)\sigma_i^2}{\sqrt{2\sum_{i=1}^{M}w_i^2(1+2\gamma_i)\sigma_i^4/N}}\right)\tag{11}$$

其中，λ_{c} 为融合中心的判决门限。采用纽曼皮尔逊准则，由式（10）推导融合中心的判决门限 λ_{c}：

$$\lambda_{\mathrm{c}}=Q^{-1}(p_{\mathrm{f}})\cdot\sqrt{2\sum_{i=1}^{M}w_i^2\sigma_i^4/N}+\sum_{i=1}^{M}w_i\sigma_i^2\tag{12}$$

3.2 MU 以概率 p_a 发送攻击

假设环境中 MU 的数量为 $J(J\leqslant M/2)$，当 MU 发送攻击时，方式有 2 种：PU 占用频段时以概率 p_a 发送“0”攻击；PU 不占用频段时以概率 p_a 发送“1”攻击，由式（7）、式（8）知 MU 发动攻击时的融合中心值为：

$$Z_{\mathrm{c}}'=\sum_{i=1}^{M}w_iY_i\pm\sum_{i=1}^{J}w_i\Delta\tag{13}$$

其中，Z_{c}' 表示 MU 发动攻击时的融合中心值。由于 $\sum_{i=1}^{M}w_i=1$ 且 $J\leqslant M/2$，可知：

$$Z_{\mathrm{c}}'=\begin{cases}\sum_{i=1}^{M}w_iY_i+\sum_{i=1}^{J}w_i\Delta, & H_0\\ \sum_{i=1}^{M}w_iY_i-\sum_{i=1}^{J}w_i\Delta, & H_1\end{cases}\tag{14}$$

同样地，Z_{c}' 也服从高斯分布，如式（15）：

$$\begin{cases}H_0:Z_{\mathrm{c}}'\sim N\left(\sum_{i=1}^{M}w_i\sigma_i^2+\sum_{i=1}^{J}w_i\Delta,2\sum_{i=1}^{M}w_i^2\sigma_i^4/N\right)\\ H_1:Z_{\mathrm{c}}'\sim N\left(\sum_{i=1}^{M}w_i(1+\gamma_i)\sigma_i^2-\sum_{i=1}^{J}w_i\Delta,2\sum_{i=1}^{M}w_i^2(1+2\gamma_i)\sigma_i^4/N\right)\end{cases}\tag{15}$$

根据式（15）知，在 H_0 情况下发送“1”攻击，$Z_{\mathrm{c}}'>Z_{\mathrm{c}}$ 需使式（12）中的门限 λ_{c} 升高得到精确的判决结果；在 H_1 情况下发送“0”攻击，$Z_{\mathrm{c}}'<Z_{\mathrm{c}}$ 需使式（12）中的门限 λ_{c} 降低得到精确的判决结果。门限 λ_{c} 需随着发起攻击时融合中心值的变化而发生移动，可变门限 λ_{c}' 为：

$$\lambda_{\mathrm{c}}'=\begin{cases}\lambda_{\mathrm{c}}+\sum_{i=1}^{J}w_i\Delta, & H_0\\ \lambda_{\mathrm{c}}-\sum_{i=1}^{J}w_i\Delta, & H_1\end{cases}\tag{16}$$

为了确定门限 λ_{c}' 具体取值，假设 Z_{H} 和 Z_{L} 分别表示融合中心统计能量的上边界值和下边界值。如图 2 所示，令 $p_{\mathrm{f}}=p(Z_{\mathrm{c}}\geqslant Z_{\mathrm{H}}\,|\,H_0)=0.05$ 或者 $p_{\mathrm{m}}=p(Z_{\mathrm{c}}\leqslant Z_{\mathrm{L}}\,|\,H_1)=0.05$，由式（10）、式（11）推导 Z_{H}、Z_{L}，得到式（17）：

$$\begin{cases}Z_{\mathrm{H}}=Q^{-1}(0.05)\cdot\sqrt{2\sum_{i=1}^{M}w_i^2\sigma_i^4/N}+\sum_{i=1}^{M}w_i\sigma_i^2\\ Z_{\mathrm{L}}=Q^{-1}(1-0.05)\cdot\sqrt{2\sum_{i=1}^{M}w_i^2(1+2\gamma_i)\sigma_i^4/N}+\\ \sum_{i=1}^{M}w_i(1+\gamma_i)\sigma_i^2\end{cases}\tag{17}$$

通过比较 Z_c 和上下边界值的大小，确定门限的 λ_{c}' 取值：

$$\lambda_{\mathrm{c}}'=\begin{cases}\lambda_{\mathrm{c}}+\sum_{i=1}^{J}w_i\Delta, & H_0,Z_{\mathrm{c}}>Z_{\mathrm{H}}\\ \lambda_{\mathrm{c}}-\sum_{i=1}^{J}w_i\Delta, & H_1,Z_{\mathrm{c}}<Z_{\mathrm{L}}\end{cases}\tag{18}$$

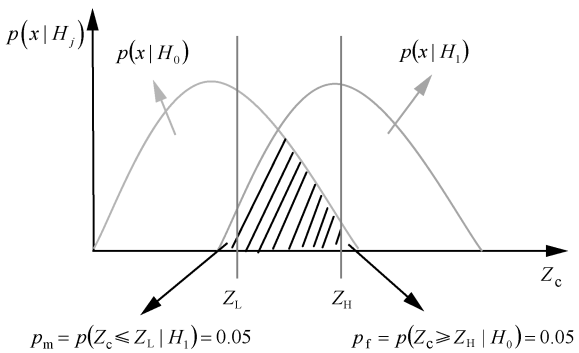


图 2 二元信号能量检测的能量值与判决概率分布

4 基于信任值的软融合方法抵御 SSDF 攻击

每个 CU 的信任值与其以往提供感知数据的行为表现有关，过去经常提供正确的感知数据使其信任值升高，错误的感知数据使信任值降低[13]。

因此，每个认知用户的信任值被看成正确感知次数与感知总数的比值。用二元假设表示第 i 个 CU 本地感知判决结果 d_i 为：

$$d_i=\begin{cases}0, & H_0\\ 1, & H_1\end{cases} \tag{19}$$

其中，“1”和“0”表示 PU 信号存在并占用频段或者不存在频段空闲的判决结果。d_{FC} 表示融合中心的判决结果，根据 CU 和 FC 所做判决结果见表 1。

表 1　K 次采样 M 个认知用户与融合中心的判决分布

采样数	CU_1	CU_2	…	CU_i	…	CU_M	FC
1	d_1	d_2	…	d_i	…	d_M	d_{FC}
2	d_1	d_2	…	d_i	…	d_M	d_{FC}
…	…	…	…	…	…	…	…
K	d_1	d_2	…	d_i	…	d_M	d_{FC}

由表 1 可知，第 i 个 CU 的信任度 R_i 为：

$$R_i=\frac{C_i}{S_i},\quad S_i\neq 0 \tag{20}$$

其中，C_i 表示在 K 次采样中的正确感知次数，S_i 是总感知次数。

很明显，当认知用户提供较大的信任值时，其在数据融合过程中就要占较大比重来改善协作频谱感知的性能。因此第 j 个 CU 分配的加权系数为：

$$w_j=R_j\Big/\sum_{i=1}^{M}R_i,\quad j=1,2,\cdots,M \tag{21}$$

总的来说，根据式（10）和式（11），由 Q 函数的性质知：p_{f} 是门限的单调递减函数，p_{m} 是门限的单调递增函数。由式（18）可知，SSDF 攻击者发送攻击时，本文提出的可变门限下的全局虚警概率 p_{f} 和漏检概率 p_{m} 的变化为：

$$\begin{cases}p_{\mathrm{f}}(\lambda_{\mathrm{c}}')<p_{\mathrm{f}}(\lambda_{\mathrm{c}}),\ H_0,\text{以概率}p_a\text{发送“1”攻击}\\ p_{\mathrm{m}}(\lambda_{\mathrm{c}}')<p_{\mathrm{m}}(\lambda_{\mathrm{c}}),\ H_1,\text{以概率}p_a\text{发送“0”攻击}\end{cases} \tag{22}$$

5　仿真结果及分析

本文给出抵御 SSDF 攻击的等增益固定门限算法以及本文提出的基于信任度的可变门限算法的仿真性能分析。考虑信道中存在加性高斯白噪声，且假设在真实的无线通信环境中，即实际存在的攻击者的所占比例少于正常认知用户所占比例。本文采用蒙特卡洛方法进行仿真，蒙特卡洛次数为 10 000，假设网络模型中存在一个 PU、一个 FC，认知用户数量 $M=10$，采样点数为 N。系统中 MU 发起攻击的概率 $p_a=0.8$，攻击强度 Δ，其攻击过程描述如下：在主用户存在的情况下以概率 $p_a=0.8$ 发送“0”攻击；主用户不存在的情况下以概率 $p_a=0.8$ 发送“1”攻击。仿真实验采用 BPSK 数字信号作为信号源，对 BPSK 进行采样，且加性高斯白噪声服从均值为零、方差为 $\sigma_i^2\left(i=1,2,\cdots,M\right)$ 的高斯分布。

图 3 给出采样数 $M=50$，信噪比 SNR=−5 dB，存在 2 个 MU，攻击概率攻击 $p_a=80\%$，强度 $\Delta=0.5$ 时，两种方案下得到的虚警概率随着给定虚警概率的变化曲线。本文提出的基于信任度的可变门限算法的全局虚警概率明显低于等增益算法。概率式 SSDF 攻击的存在，导致 CU 发送至融合中心的感知数据的能量值发生较大偏移，融合中心进行数据融合得到相对较高的 p_{f}。因为在 PU 不占用频段，MU 发动攻击导致融合中心的能量统计值升高，本文算法通过提高门限降低系统的虚警概率 p_{f}。图 3 给出的攻击强度为 0.5，相对较小，当给定虚警概率 $p_{\mathrm{f}}=0.1$，两种算法的虚警概率依次为 0.3、0.06。相比较得到，本文算法得到较低的虚警概率。

图 4 是采样数 $M=50$，信噪比 SNR=−5 dB，存在 2 个 MU，攻击概率攻击 $p_{\mathrm{a}}=80\%$，强度 $\Delta=0.5$、1、2 时，本文提出可变门限算法的全局虚警概率 p_{f} 随着给定虚警概率的变化曲线。从图 4 可以看出，随着攻击强度的增加，全局虚警概率 p_{f} 随之增加，但增幅较小，即使是在攻击强度 $\Delta=2$ 的强攻击下，本文算法依然可以得到较低的虚警概率。

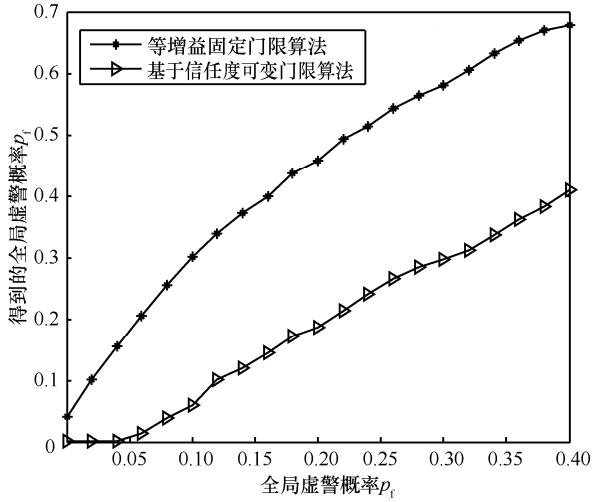


图 3　不同方案下随 p_f 变化的虚警概率曲线

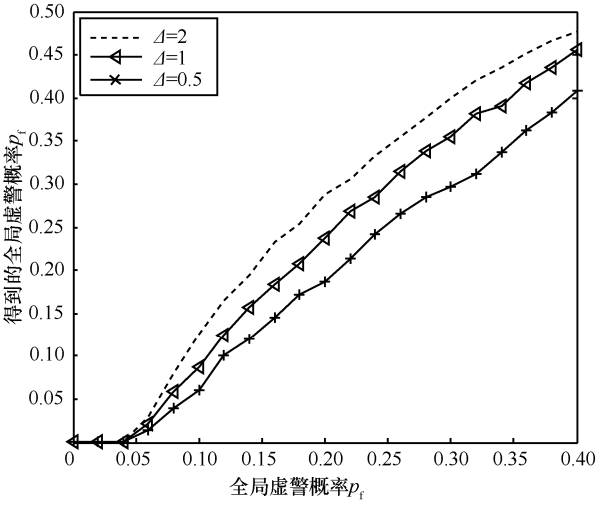


图 4　不同攻击强度下基于信任度的可变门限算法随 p_f 变化的虚警概率曲线

图 5 描述了采样数 $M=50$，信噪比 SNR = −5 dB，存在 2 个 MU，攻击概率攻击 $p_a=80\%$，强度 $\Delta=2$ 时，两种算法的全局漏检概率 p_m 随给定虚警概率的变化曲线。由图 5 可知，两种算法的漏检概率都随着虚警概率的增加而降低，同种条件下可变门限算法比等增益固定门限算法的漏检概率降低较快，尤其是虚警概率 p_f 较低的情况下，在 $p_f=0.1$，两种算法的漏检概率分别为 0.072 5、0.008 5。其主要原因是 PU 工作的情况下，攻击者发送恶意攻击导致融合中心的统计量变小，本文算法通过降低可变门限值导致判决为“0”的情况有所减少进而得到较低的漏检概率 p_m。

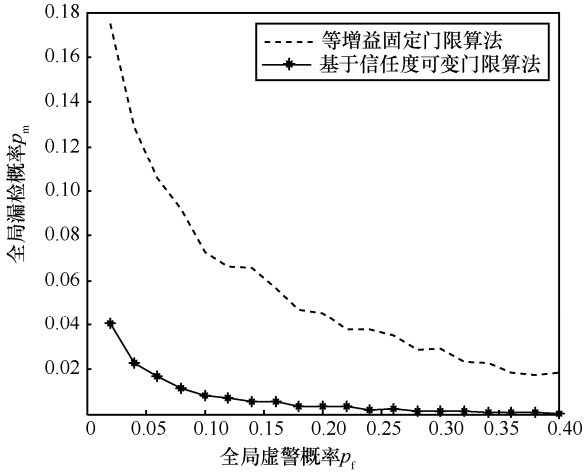


图 5　不同方案下随 p_f 变化的漏检概率曲线

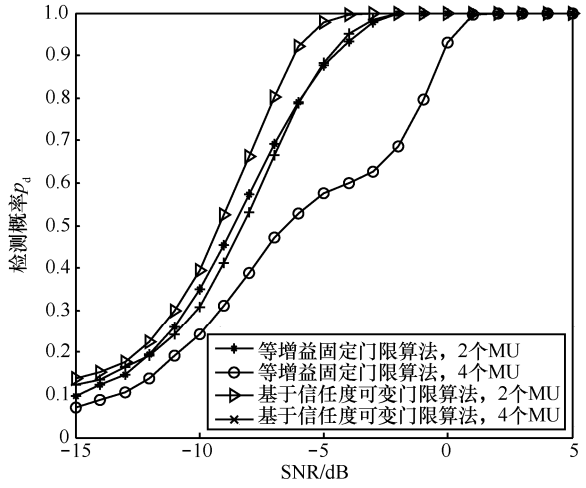


图 6　不同方案下随信噪比变化的检测概率 p_d 的曲线

图 6 给出采样数 $M=50$，攻击概率攻击 $p_a=80\%$，攻击强度为最大攻击强度即信号与噪声的总能量且 $p_f=0.05$，MU 数量分别为 2 个和 4 个时，两种算法系统检测概率 p_d 随信噪比的变化。从图 6 知，随着信噪比的增加两种算法的检测概率都上升，在攻击者数量一样的情况下本文优化算法比等增益固定门限算法的检测概率要高且能较快达到 1。在 SNR=−10 dB，攻击者数量为 4 时，两种算法的检测概率分别为 0.241 7、0.307 5，检测性能提高了 27%。

6　结束语

本文针对概率式 SSDF 攻击，提出基于信任

度的可变门限能量检测协作频谱感知算法。本算法的门限值不再是固定不变的，而是随着恶意用户发动攻击的情况发生变化，其过程是通过比较融合中心真实值与上下边界值的大小决定门限增大或减小，然后将实际融合中心值与变化过的门限进行比较，得到最终判决。算法中的信任度主要依据以往提供感知数据的行为表现进行更新，将正确感知次数与总感知次数的比值看成信任度，合理分配权重。仿真结果表明，该算法与传统算法相比，能有效抵御 SSDF 攻击，在降低漏检概率和虚警概率的同时提高了系统的检测概率。

参考文献：

[1] YANG J X, CHEN Y B, SHI W G, et al. Cooperative spectrum sensing against attacks in cognitive radio networks [C]// 2014 IEEE International Conference on Information and Automation (ICIA), July 26-31, 2014, Hulun Buir, China. Piscataway: IEEE Press, 2014: 71-75.

[2] 冯景瑜, 李金龙, 卢光跃. 协作频谱感知中抗 SSDF 攻击的认知用户不确定性行为评估[J]. 电信科学, 2015, 31(2): 97-102.
FENG J Y, LI J L, LU G Y. Evaluating uncertainty behaviors of cognitive users against SSDF attack for cooperative spectrum sensing[J]. Telecommunications Science, 2015, 31(2): 97-102.

[3] 裴庆祺, 李红宁, 赵弘洋. 认知无线电网络安全综述[J]. 通信学报, 2013, 34(1): 144-158.
PEI Q Q, LI H N, ZHAO H Y, et al. Security in cognitive radio networks[J]. Journal of communication, 2013, 34(1): 144-158.

[4] XIAO J, CHEN Y B, XING C X, et al. An optimized scheme to resist primary user emulation attacks[C]// The 2016 International Conference on Communications, Information Management and Network Security(CIMNS2016), Sept 25-26, 2016, Shanghai, China. [S.l.:s.n.], 2016: 176-180.

[5] 卢光跃, 苏杭. 分布式协作认知无线电 SSDF 攻击的防御策略综述[J]. 电信科学, 2017, 33(1): 95-105.
LU G Y, SU H. Survey on SSDF attack and defense for distributed cooperative cognitive radio[J]. Telecommunications Science, 2017, 33(1): 95-105.

[6] 冯景瑜. 协作频谱感知中的 SSDF 攻击及其对策研究[J]. 电信科学, 2014, 1(11): 67-72.
FENG J Y. Research on SSDF attack and defense for cooperative spectrum sensing[J]. Telecommunications Science, 2014, 1(11):67-72.

[7] RAWAT A S, ANAND P, CHEN H, et al. Collaborative spectrum sensing in the presence of byzantine attacks in cognitive radio networks[J]. IEEE Transactions on Signal Processing, 2011, 59(2): 774-786.

[8] ZHAO T, ZHAO Y. A new cooperative detection technique with malicious user suppression[C]// IEEE ICC2009, June 14-18, 2009, Dresden, Germany. Piscataway: IEEE Press, 2009: 1-5.

[9] SHARIFI A A, NIYA M J M. Defense against SSDF attack in cognitive radio networks: attack-aware collaborative spectrum sensing approach[J]. IEEE Communications Letters, 2016, 20(1): 93-96.

[10] PENG T, CHEN Y B, XIAO J, et al. Improved soft fusion-based cooperative spectrum sensing defense against SSDF attacks[C]// IEEE International Conference on Computer Information and Telecommunication System(CITS 2016), July 6-8, 2016, Kunming, China. Piscataway: IEEE Press, 2016: 134-138.

[11] YANG Z L, CHEN Y B, XING C X, et al. Improved reputation of soft fusion based on cooperative spectrum sensing defence SSDF attacks[C]//The 2016 International Conference on Communications, Information Management and Network Security(CIMNS2016), Oct 22-23, 2016, Beijing, China. [S.l.:s.n.], 2016: 181-185.

[12] ZHOU M, SHEN J, CHEN H, et al. A cooperative spectrum sensing strategy based on the Bayesian reputation model in cognitive radio networks[C]//2013 IEEE Wireless Communication and Networking Conference(WCNC 2013), April 7-10, 2013, Shanghai, China. Piscataway: IEEE Press, 2013: 614-619.

[13] FENG J Y, ZHANG Y Q, LU G Y, et al. Defend against collusive SSDF attack using trust in cooperative spectrum sensing environment[C]// IEEE International Conference on Trust, Security and Privacy in Computing and Communication, Oct 19-20, 2014, Beijing, China. Piscataway: IEEE Press, 2014: 1656-1661.

[作者简介]

肖洁（1992−），女，云南民族大学硕士生，主要研究方向为认知无线电网络安全、频谱感知数据篡改（SSDF）攻击。

陈跃斌（1963−），男，云南民族大学教授、硕士生导师，主要研究方向为认知无线电、协作频谱感知和认知无线电网络安全。

陈楚天（1991−），男，云南民族大学硕士生，主要研究方向为认知无线电网络安全、频谱检测。

郑婷（1992−），女，云南民族大学硕士生，主要研究方向为认知无线电网络安全、恶意模拟主用户（PUEA）攻击。

钱继武（1990−），男，云南民族大学硕士生，主要研究方向为认知无线电网络安全、频谱感知。

综述

非正交多址系统资源分配研究综述

王正强，成蕖，樊自甫，万晓榆
（重庆邮电大学，重庆 400065）

摘　要：非正交多址（NOMA）是 5G 无线网络的一个重要候选技术，可以满足下一代移动通信系统低时延、低功耗、高可靠、高吞吐量、广覆盖等需求。NOMA 通过在发送端采用叠加编码和接收端采用串行干扰消除来实现在同一资源块复用多个用户数据，从而相对于传统的正交多址接入方式提高了频谱效率。概述了 NOMA 系统资源分配的研究现状，其中包括单载波 NOMA 的资源分配、多载波 NOMA 的资源分配、协作 NOMA 中继的资源分配、硬件损伤条件下协作 NOMA 的资源分配。最后，总结了当前研究中存在的主要问题，讨论了 NOMA 资源分配技术的研究挑战和未来研究方向。

关键词：非正交多址；资源分配；单载波；多载波；中继；硬件损伤
中图分类号：TN925
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018236

A survey of resource allocation in non-orthogonal multiple access systems

WANG Zhengqiang, CHENG Qu, FAN Zifu, WAN Xiaoyu
Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: Non-orthogonal multiple access (NOMA) is an important candidate technology of the fifth generation (5G) wireless network, which can meet the low latency, low power consumption, high reliability, high throughput, wide coverage requirements of the next generation mobile communication systems. NOMA utilizes superposition coding at the transmitter and successive interference cancellation at the receiver to allow several users' data multiplexing in the same resource block. It improves the spectrum efficiency compared with the traditional orthogonal multiple access. The research status of resource allocation of NOMA systems was summarized including single-carrier NOMA resource allocation, multi-carrier NOMA resource allocation, cooperative NOMA relay resource allocation, and cooperative NOMA resource allocation under hardware impairment conditions. Finally, the main problems in the current study were summarized and the research challenges and some future research directions of NOMA resource allocation technology were discussed.

Key words: non-orthogonal multiple access, resource allocation, single-carrier, multi-carrier, relay, hardware impairment

收稿日期：2018−02−03；**修回日期**：2018−07−13
基金项目：国家自然科学基金资助项目（No.61701064）；重庆市教委科学技术项目（No.KJ1600424）
Foundation Items：The National Natural Science Foundation of China (No.61701064), Scientificand Technological Research Program of Chongqing Municipal Education Commission (No.KJ1600424)

1　引言

随着移动互联网、社交网络和物联网的飞速发展，移动智能终端日益普及，爆炸式增长的移动数据业务对无线通信系统的要求越来越高；同时，爆发式增长的数据流量给有限的频谱资源也带来了巨大的挑战[1-2]。因此，能够支持更多用户连接、更高频谱效率和能量效率的新型多址接入技术成为 5G 产业界和学术界一个迫切需要解决的难题和研究热点[3]。在最新的 5G 新型多址技术研究中，基于功率域复用的非正交多址接入（non-orthogonal multiple access，NOMA）技术是 5G 网络的一个重要的候选技术，日益受到产业界的重视，不仅可满足 5G 在频谱效率和连接数等方面的需求，还可以满足低时延、高可靠性、大规模连接、提高公平性和高吞吐量的异构需求[4-6]。与传统的正交多址接入（orthogonal multiple access，OMA）方案不同，NOMA 可以利用功率域、时域、频域、码域来实现多路访问，在发送机端采用非正交发送，主动引入干扰信息，在接收端通过串行干扰消除（successive interference cancellation，SIC）技术按照一定的顺序进行多用户检测、正确解调以及干扰消除，以便获得自己的信息[7]。近年来，在工业界和学术界已经研究了许多NOMA技术，例如交织分多址接入（interleave division multiple access，IDMA）、比特分割多路复用（bit division multiplexing，BDM）、稀疏码多址接入（sparse code multiple access，SCMA）、多用户共享访问接入（multi-user sharing access，MUSA）和模式分割多址（pattern division multiple access，PDMA）[8]。NOMA 技术相对于传统的 OMA 技术具有如下优点：更高的频谱效率、更高的小区边缘吞吐量、更低的传输等待时间、增强的用户公平性和更多的用户连接数[2]等。基于 NOMA 技术的以上优点，尤其是满足未来宽带无线通信技术的高频谱效率和大连接需求，已成为当前研究的一项关键技术和热点。用户功率域的非正交性使得资源分配算法在 NOMA 系统中变得尤为重要，只有通过有效的资源分配算法才能保证 NOMA 系统接收端的用户数据被有效解码，实现更大的用户连接数、更高的频谱效率和能量效率。考虑 NOMA 系统采用的载波数目和是否采用协作通信技术，当前研究中的 NOMA 系统分配问题主要可以分为基于单载波 NOMA[9-22]、基于多载波 NOMA[23-41]、基于协作 NOMA[42-54]和硬件损伤条件下协作 NOMA[55-57]资源分配 4 个方面。

2　NOMA 资源分配

2.1　单载波 NOMA 资源分配

当 NOMA 原理被应用于单个正交资源块即单载波时，实现多址接入的频谱有效方式是利用功率域，大部分工作是在图 1 的系统模型下进行研究的。参考文献[9]考虑只有两个用户的下行 NOMA 系统，在基站总功率和用户最小速率需求的约束条件下，研究了系统的和速率最大化问题。基于优化问题的 KKT（Karush-Kuhn-Tucker）条件，得出带有闭式解的最优功率分配方案。参考文献[10]在固定功率分配下，研究单载波 NOMA 系统的用户配对对于系统性能的影响，证明了当配对用户的信道增益差异越大时，采用固定功率分配的 NOMA 系统的和速率比 OMA 方式的和速率越高。参考文献[11]考虑 MIMO-NOMA 分层传输系统中的最佳功率分配，在复用两个用户情况下提出一种分层传输的 MIMO-NOMA 总速率最大化算法。参考文献[12]研究 MIMO-NOMA 系统的遍历容量最大化问题。首先得出最优的信道输入协方差矩阵，然后提出最优功率分配方案和低复杂度的次优功率分配方案。因为无法获得功率分配的闭式表达式，因此在总传输功率限制和最小速率约束条件下，采用二分搜索法来获得用户 2 的功率分配，最大化系统的遍历容量。而后，为了降低复杂度，由于优化问题的解位于可行区域的边界上，还提

出了近似最优功率分配算法。结果表明，提出的NOMA 方案明显优于传统正交多址方案。参考文献[13]基于用户分簇研究了 MIMO-NOMA 系统的公平性，研究了动态用户分配和功率优化问题。由于将用户分配到不同簇是一个 NP 问题，因此，本文提出了次优算法。与穷举搜索方法相比，所提算法在吞吐量和复杂性之间实现了很好的折中。参考文献[14]基于比例公平调度准则，研究了两个用户下行 NOMA 系统，提出了最大系统和速率与最大最小用户速率的两种功率分配算法。参考文献[15]研究基于 NOMA 的无线能量传输网络中的资源分配问题，给出了联合优化基站发射功率、能量采集和信息传输时间的算法以最大化系统速率。仿真结果表明与固定发射功率的 NOMA 方案相比，所提方案可以实现更高的速率与系统公平性。参考文献[16]在总发射功率和用户最小速率约束下，研究衰落 MIMO-NOMA 系统的能效优化问题，提出了近似最优功率分配方案，并获得了次优闭式解。仿真结果表明所提 NOMA 方案在频谱效率和能量效率方面优于传统的 OMA 方案。参考文献[17]在多用户下行系统中从能效角度研究了基于认知无线电技术启发的 NOMA 系统。在每个主用户受服务质量约束的条件下，最大化系统的能效。仿真结果表明 NOMA 比传统的 OMA 能效更高。参考文献[18]基于单输入单输出（single input single output，SISO）NOMA 下行系统，在用户最小速率需求约束下，提出一种内外两层迭代算法来最大化系统能效。其中内层算法在固定系统能效的情况下对用户进行功率分配，外层算法采用二分法搜索能效。仿真结果表明 NOMA 系统获得的能效优于传统 OMA 方案。参考文献[19]推导了理想信道下单载波 NOMA 系统中基于比例公平调度的最优功率分配的闭式解，并联合功率分配和用户集选择设计了一种低复杂度算法。结果表明，所提出方案所获得的吞吐量接近上限，且相比传统 OMA、分式发射功率分配（fractional transmit power allocation，FTPA）等方案性能最优。参考文献[20]采用 Stackelberg 博弈研究了 NOMA 系统基于定价的功率分配问题来最大化基站的收益，通过引入辅助变量将基站优化问题转化为 3 个优化子问题，进一步利用凸优化和交替优化来分配用户功率，仿真结果表明所提算法比等功率分配算法提高了基站收益。参考文献[21]针对 NOMA 系统提出了一种新的基于定价的功率分配算法。首先，针对两个用户的情况给出闭式解，分析表明基站的最优策略为向两个用户或仅向信道较好的用户分配功率；然后，针对用户数目多余两个的情况，提出了基于两用户轮询配对的迭代算法。仿真结果表明所提出算法在基站的收益和用户的和速率方面更优[20]。考虑用户之间的速率公平比例约束条件，参考文献[22]采用 Stackelberg 博弈研究了 NOMA 系统基于定价机制下基站收益最大化问题。首先通过用户功率和速率之间的对应关系，将基站的功率分配问题转化为速率优化问题；进一步利用速率优化问题的单调性，提出了在速率公平比例约束条件下 NOMA 系统基于定价的最优功率分配算法。

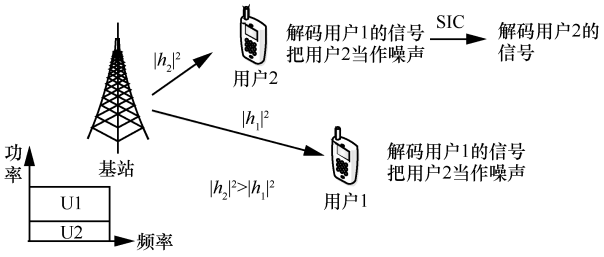


图 1　两用户下行链路单载波 NOMA 方案示例

2.2　多载波 NOMA 的资源分配

由于资源分配的灵活性和多用户分集增益，多载波技术[23]已在过去的宽带无线通信中被大量采用。在传统的多载波系统中，给定的频谱资源被分成多个子载波，而每个子载波至多分配给一个用户使用，有效避免多用户干扰。然而这样的分配方式却不能达到容量区域的上界，特别是当系统保证用户公平性时，会将部分子载波专门分

配给信道质量较差的用户使用，从而会造成频谱资源的浪费[24-25]。因此，联合多载波技术和 NOMA 技术的多载波 NOMA 接入方式可以进一步提高频谱利用效率，这已经成了未来宽带无线通信接入技术的一个演进方向。对于多载波的资源分配，参考文献[26]在图 2 的系统模型下，通过联合子载波分配和功率分配优化多载波 NOMA 系统的加权和速率。因为优化问题是 NP 问题，所以将子信道分配问题转化为多对多的用户与子信道双向匹配问题，提出用户—子信道交互匹配算法（user- subchannel swap-matching algorithm，USMA-1）与 USMA-2 两种算法，将功率分配问题转化为几何规划问题，最后采用联合子信道和功率分配算法（joint subchannel and power allocation algorithm，JSPA）。仿真结果表明所提算法在用户和速率和公平性方面都优于传统的 OFDMA（orthogonal frequency division multiple access）方案。参考文献[27]研究两个用户多载波 NOMA 系统，在用户传输速率和最大发送功率限制条件下，最小化子载波数，仿真结果表明，所提出的资源分配策略提高了频谱效率和小区边缘用户吞吐量。参考文献[28]研究了在每个子载波最多复用两个用户的情况下，多载波 NOMA 系统的加权系统吞吐量最大化问题。基于单调优化和凸逼近的方法，分别提出最优方案和低复杂度的次优算法。仿真结果表明所提次优算法性能相对于传统的多载波正交多址系统提高了系统的吞吐量。参考文献[29]研究了多载波 NOMA 系统最大化和速率的资源分配问题，由于是非凸优化问题，因此，提出采用匹配理论和注水功率分配的次优算法进行求解，仿真结果表明所提算法获得的系统速率优于基于 OFDMA 的分配方式。参考文献[30-31]考虑下行 NOMA 系统，联合优化信道和功率分配以最大化系统加权和速率问题。由于该优化问题是 NP 问题，因此提出基于拉格朗日对偶和动态规划的次优算法。仿真结果表明所提算法优于 OFDMA 和带有分数功率分配的 NOMA 方案[32]。参考文献[33]考虑多载波 NOMA 系统，提出一种两次迭代的注水算法，并证明了该算法在每个子载波复用不超过两个用户的情况下具有收敛性，仿真结果表明所提算法性能接近现有的次优算法[30]，但具有更低的时间复杂度。参考文献[34]在满足用户最小速率需求的情况下，研究了多载波 NOMA 系统总功率最小化问题，并提出了低复杂度的联合子载波和功率分配算法。仿真结果表明所提算法相比传统 OFDM 的频分复用和静态的非正交资源分配算法降低了系统的能量消耗。参考文献[35]考虑在用户最小速率约束条件下，研究多载波 NOMA 系统和速率的最大化问题。由于该问题是非凸的，因此，本文提出次优算法将原问题分解为子载波分配和功率分配问题。先在假设等功率分配的情况下进行子载波分配，然后在给定的子载波分配的情况下进行功率分配。参考文献[36]在发送端具有信道状态统计信息（channel state information at the transmitter，CSIT）的前提下，提出一种次优的功率分配和用户调度算法来得到多载波 NOMA 系统的最小发射总功率。仿真结果表明，所提算法相比传统 OMA 方式降低了系统的总功率。参考文献[37]考虑基站全双工多载波 NOMA 系统的资源分配问题，优化系统的加权和速率，提出利用连续凸逼近的次优算法来平衡算法复杂度和最优性。结果表明所提算法接近最优性能，并且在系统平均吞吐量、平均接入用户数、系统公平性上优于 3 种基准的对比算法（基于全双工的多载波正交接入、半双工的多载波 NOMA 和半双工的多载波 OMA）。参考文献[38-39]在每个子载波复用最多两个用户的限制条件下，研究多载波 NOMA 系统基于能效的子信道分配和功率分配算法，提出了一种次优的信道分配和功率分配算法以最大化 NOMA 系统用户的能效的总和。相比传统的 OFDM 方案，所提 NOMA 方案能实现更好的和速率与能量效率。参考文献[40]研究了由多载波—非

正交多址支持的虚拟无线网络（virtualized wireless network，VWN）的上行链路资源分配问题，将优化问题分解为独立的功率和子载波分配问题，并提出一种基于连续凸近似和互补几何规划的迭代算法。结果表明，与 OMA 相比，所提出的多载波 NOMA 算法可以显著提高频谱和功率效率。参考文献[41]通过共同考虑信道分配和功率控制，为基于 NOMA 的上行链路网络制定了一个和速率最大化问题，将原始问题转化为图论中的最大加权独立集问题，提出一种有效的低复杂度资源分配算法。结果显示该算法在数据速率和支持用户数方面相对其他方案的性能更优。

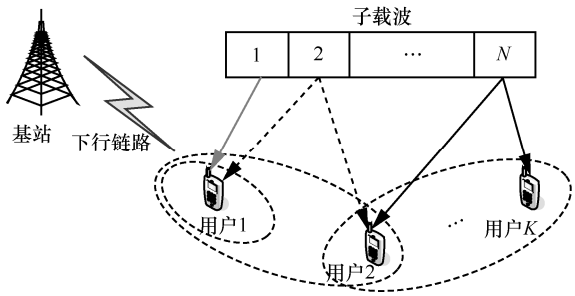


图 2　多用户下行链路多载波 NOMA 方案示例

2.3　协作 NOMA

协作通信可以使用多个中继来帮助源节点与目的节点通信，以提高无线网络的容量和可靠性，也是防止无线信道多径衰落和提高系统吞吐量性能的最有效方法之一，还具有降低发射功率和减少能量消耗的特点[42]。又由于受到 NOMA 优势的吸引，近年来部分研究人员将 NOMA 的技术和中继技术相结合，开展了协作 NOMA 中继系统的性能分析[43-49]与资源分配[50-54]研究。

2.3.1　性能分析

参考文献[43]在协调直接和中继传输（coordinated direct and relay transmission，CDRT）中引入 NOMA，导出了中断概率和遍历和容量的分析表达式，仿真结果表明与非协调直接和中继传输中的 NOMA 系统相比，所提方案具有显著的性能增益。参考文献[44]研究了在不理想信道状态信息的 Nakagami-*m* 衰落信道下，协作 NOMA 下行放大转发中继网络的中断概率，仿真结果表明相比传统 OMA 系统，性能可以显著提升。参考文献[45]研究了 NOMA 多天线中继网络中用户的中断行为，仿真结果表明当中继位置靠近移动用户时，OMA 可实现更好的中断性能，但 NOMA 可提供更好的频谱效率和用户公平性。参考文献[46]在 AF 中继的帮助下研究了协作动态 NOMA 网络的中断性能，导出了中断概率精确闭合表达式的近似结果，仿真结果表明协作 NOMA 比协作 OMA 有更高的分集增益和编码增益，中断性能也有所提高。参考文献[47]分析并比较了 NOMA 协作和 NOMA 时分多址两种方案的中断性能，结果表明 NOMA 协作方案的中断性能比 NOMA 时分多址方案更好。参考文献[48]研究了具有多个中继的协作下行链路非正交多址网络的中继选择方案。提出了两个最优中继选择方案，称为两阶段加权最大最小值（weighted-max-min，WMM）和最大加权谐波均值（maxweighted-harmonic-mean，MWHM）方案。分析两种方案的中断概率，并确定它们的分集增益。结果表明，所提出的最优两阶段 WMM 和 MWHM 方案优于现有的次优中继选择（relay selection，RS）方案。参考文献[49]研究了中继选择对协作 NOMA 性能的影响，其中继以全双工（full-duplex，FD）或半双工（half-duplex，HD）模式工作，并采用随机几何对网络的中继位置进行建模，推导出 FD / HD NOMA 两种 RS 方案的中断概率解析表达式。结果表明，基于 FD 的 RS 方案在低信噪比（SNR）区域中具有比基于 HD 的 RS 方案更好的中断性能；基于 FD/HD 的 NOMA 中继选择方案（single-stage RS，SRS）/两阶段中继选择方案（two-stage RS，TRS）的中断行为优于随机中继选择方案（random RS，RRS）和基于 OMA 的中继选择方案。

2.3.2　资源分配

参考文献[50]分析了协作 NOMA 中继系统的平均速率的渐进表达式，提出 NOMA 的次优功率分配方案，仿真结果表明此系统能提高频谱效率。参考文献[51]研究 NOMA 协作中继系统（cooperative relaying system using non-orthogonal multiple access，CRS-NOMA）的新型检测方案，目的节点通过采用最大比合并与串行干扰消除直连信号和转发信号进行联合解码，研究了系统的遍历和速率和中断性能，通过求解遍历和速率对于功率分配因子的导数，可得到最优功率分配因子即所提功率分配方案。所提方案比参考文献[50]方案在遍历和速率和中断性能方面都有明显的改善。参考文献[52]研究协作非正交多址中继（collaborative noma assisted relaying，CNAR）系统，分析 CNAR 系统和简化的 CNAR 系统中断行为，分别考虑源—中继和中继—目的节点链路的中断行为来分析两个系统的中断概率，提出通过最小化中断概率来保证数据速率的最优功率分配方案。结果证明了所提出的 CNAR 在可能的传输策略中实现了最佳性能，并且简化的 CNAR 获得了类似的性能并且降低了中继复杂性。参考文献[53]在图 3 的系统模型下，研究了具有单向 OFDM 放大转发中继的 NOMA 系统的资源分配问题，通过优化子信道分配和功率分配来最大化平均和速率。将问题转化为多对多双向匹配问题，提出两个近似最优的源—目的节点间的子信道匹配算法，即静态匹配算法和动态匹配算法都在有限次数的迭代之后收敛到成对的稳定匹配，再采用注水功率分配算法进行功率分配。结果表明所提算法具有较低复杂度，能服务更多用户且平均和速率也高于传统 OFDMA 系统。参考文献[54]提出一种基于 NOMA 的新型协作传输方案来重新设计无线回程双层异构网络架构。设计 NOMA 解码顺序以及在宏基站（macro base station，MBS）和小小区接入点（small cells access point，SCAP）处的下行链路发射波束成形和功率分配，最大化可达速率和满意的用户数。提出了一种基于连续凸近似和主要最小化方法的迭代低复杂度算法来求次优解。结果表明，所提方案更加先进和有效，且总可达速率方面优于常规设计。

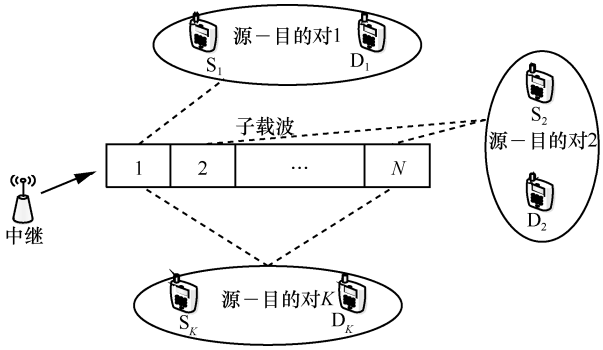


图 3　多用户协作 NOMA 中继方案示例

2.4　硬件损伤条件下协作 NOMA 的资源分配

在当前已有研究中，大多研究和文献都是在理想的硬件条件假设下进行的分析和优化。然而，在实际的协作 NOMA 系统中，硬件并不都是完美的，硬件遭受来自各个方面的影响和损害，源节点、中继节点以及用户节点存在不同程度的硬件损伤，这些硬件损伤包括：把通信设备自身产生并难以消除的系列有损害通信质量的诸多因素，包括 I/Q 不平衡、非线性功放产生的等效噪声以及射频电路噪声等。一般来说，通过在发射机处使用某些校准技术或/和在接收机处的补偿算法，通常可以减轻由上述单一类型的硬件损伤导致的性能退化。然而，这些方法不能完全消除硬件损伤，因此总是存在一定量的由于残留硬件损伤（residual hardware impairment，RHI）而导致的未被计入的失真，这些损失被添加到发送/接收信号中，导致系统性能的下降。参考文献[55]研究了量化残余硬件损伤（RHI）对基于非正交多址（NOMA）的中继网络的影响，推导出中断概率的精确和渐近表达式，在 Nakagami-*m* 衰落信道上给出封闭形式。研究结果表明，在低 SNR 或低目标速率下，RHI 引起的中断性能损失较小，但在高

SNR 或目标速率下损失显著。此外，还提出了系统遍历和速率（ergodic sum rate，ESR）的渐近表达式，并与正交多址（OMA）传输的传统硬件损坏中继系统的 ESR 作对比，结果表明，在没有 RHI 的情况下，NOMA 或 OMA 系统中的 ESR 随着 SNR 的增加而单调增加，而在两个系统的硬件受损情形中引入了不可避免的 ESR 上限。参考文献[56]量化联合发射机/接收机同相正交相位不平衡（in-phase/quadrature-phase imbalance，IQI）所带来的硬件损伤对多径衰落条件下基于 NOMA 的多载波系统性能的影响，并推导了所考虑的多载波 NOMA 建立的渐近分集阶数。结果证明同相正交相位不平衡 IQI 的影响在 NOMA 用户中差异很大，并取决于底层的系统参数。参考文献[57]研究了非正交多址（NOMA）双跳（dual-hop，DH）放大转发中继网络的性能，其中考虑 Nakagami-*m* 衰落信道。综合考虑源节点、中继节点和目的节点处的收发器硬件损伤，导出了中断概率和近似遍历和速率的闭式表达式。另外，为了进一步揭示硬件损伤参数对网络性能的影响，对高信噪比（SNR）情况下的中断概率和遍历和速率进行了渐近分析。

2.5 对比与分析

对以上关于 NOMA 资源分配算法相关工作进行总结见表 1。

3 研究挑战和未来研究方向

尽管目前国内外对于 NOMA 系统的资源分配问题的研究已经取得了一定的研究成果。上述提到的资源分配方法也在一定程度上满足了 NOMA 系统某一方面的优化目标，但是现有研究针对 NOMA 系统的资源分析手段和优化方法仍然具有一定的局限性。由以上归纳分析发现，大部分研究都是理想信道状态信息下，进行算法设计和优化，但无线信道状态信息通常会受到信道估计误差、量化误差等实际因素的影响，因此需要进一步考虑非理想信道状态信息下，对于单载波 NOMA、多载波 NOMA、协作 NOMA 的资源分配问题进行建模和分析，又由于该问题一般为非凸甚至 NP 难问题。因此，如何设计出低复杂度的在线资源分配算法，接近最优性能的次优算法也极其重要和具有挑战。

基于 Stackelberg 博弈方法采用定价机制研究 NOMA 系统的资源分配问题[20-22]仍然处于起

表 1 NOMA 资源分配相关研究工作总结

	研究目标	参考文献	解决方法
单载波 NOMA	和速率	[9-15,19]	凸优化、KKT 条件、穷举搜索算法、两层迭代算法、二分搜索法、变量替换、交替优化、博弈论
	公平性	[13-15]	
	能量效率	[16-18]	
	基站收益	[20-22]	
多载波 NOMA	和速率	[26-31,35,37,41]	几何规划、迭代算法、连续凸逼近、单调优化、凸优化、匹配理论、注水功率分配算法、对偶分解、拉格朗日对偶法、动态规划
	公平性	[33,37]	
	功率	[34,36,40]	
	能量效率	[38-39]	
协作 NOMA 中继	和速率	[43,50-54]	凸优化、匹配算法、分式规划、单调优化
	中断性能	[43-49,52]	
	公平性	[45]	
硬件损伤下协作 NOMA	和速率	[55,57]	
	公平性	[55]	
	中断性能	[55-57]	

步阶段，在未来研究中可以进一步考虑用户的服务质量需求，联合容许控制和功率分配设计 NOMA 系统基于定价的资源算法来最大化基站的收益。

现有研究针对 NOMA 系统在硬件损伤下的性能分析也刚处于研究起步阶段，相关研究成果主要有参考文献[55-57]，这些研究主要分析硬件损伤下的 NOMA 系统的中断概率和渐进和速率等性能指标，但并没有从资源分配角度对系统资源进行优化。因此，可以对硬件损伤条件下协作 NOMA 系统的基于频效的资源分配研究，充分考虑发射源端、中继节点、用户接收端的硬件损伤程度，信道的状态信息、发射功率等因素的影响，利用连续凸逼近、拉格朗日对偶分解等优化理论来设计；也可以对硬件损伤条件下协作 NOMA 系统的基于能效的资源分配研究，综合利用分式规划理论、单调优化方法等建立基于硬件损伤条件下的协作 NOMA 系统基于能效的低复杂度资源分配算法；还可以对硬件损伤条件下协作 NOMA 基于经济效益的资源分配研究，采用 Stackelberg 博弈建模分析协作 NOMA 系统中继节点在价格激励机制下的最优定价和功率分配策略，最大化中继节点和用户的收益。

综上所述，对于单载波 NOMA[9-22]、多载波 NOMA[26-41]已有的研究成果大多集中在和速率、公平性、功率、能效等方面；对于协作 NOMA[43-54]已有的研究成果大多集中在中断概率、和速率和公平性等方面；对于硬件损伤条件下协作 NOMA[55-57]的已有研究成果也大多集中在和速率、公平性、中断性能等方面。因此，针对这几个方面，研究和设计基于其他性能指标的有效资源分配算法也是极其重要的，可以采用的合理相关数学工具有：分式规划、连续凸逼近、博弈论、压缩不动点、对偶分解、顽健优化、随机过程、动态规划等方法。并通过计算机仿真和实验去验证评价算法性能。进而设计出低复杂度、顽健的优化算法，实现系统的绿色节能。

4　结束语

本文先简述了 NOMA 的原理与优势，然后基于单载波 NOMA、多载波 NOMA、协作 NOMA 中继、硬件损伤条件下协作 NOMA 的资源分配进行介绍，最后总结了当前现状，并提出研究挑战和未来发展方向。但是现有研究对 NOMA 资源分配研究考虑的模型大部分过于理想，大都假设系统具有理想信道状态信息，用户的接收端可以完全消除信道条件较好用户的干扰。然而，考虑到实际通信系统信道估计误差、信道反馈误差、量化误差等因素的影响和用户解码硬件的限制，有必要在信道估计误差和不完全干扰消除情况下，对 NOMA 系统的资源分配问题进行建模和分析，进一步研究和探索具有顽健性的资源分配算法。

参考文献：

[1] LI Q C, NIU H N，PAPATHANASSIOU A T, et al. 5G network capacity: key elements and technologies[J]. IEEE Vehicular Technology Magazine, 2014, 9(1): 71-78.

[2] DAI L L, WANG B C, YUAN Y F, et al. Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends[J]. IEEE Communications Magazine, 2015, 53(9): 74-81.

[3] TAO Y Z, LIU L, LIU S, et al. A survey: several technologies of non-orthogonal transmission for 5G[J]. China Communications, 2015, 12(10): 1-15.

[4] 毕奇,梁林,杨姗,等. 面向5G的非正交多址接入技术[J]. 电信科学,2015, 31(5): 20-27.
BI Q, LIANG L, YANG S, et al. Non-orthogonal multiple access technology for 5G systems[J]. Telecommunications Science, 2015, 31(5): 20-27.

[5] DING Z G, LIU Y W, CHOI J, et al. Application of non-orthogonal multiple access in LTE and 5G networks[J]. IEEE Communications Magazine, 2017, 55(2): 185-191.

[6] WEI Z Q, YUAN J H, NG D W K, et al. A survey of downlink

non-orthogonal multiple access 5G wireless communication networks[J]. ZTE Communication, 2016, 14(4): 17-25.

[7] DING Z G, PENG M G, POOR H V. Cooperative non-orthogonal multiple access in 5G systems[J]. IEEE Communications Letters, 2015, 19(8): 1462-1465.

[8] CHEN S, REN B, GAO Q, et al. Pattern division multiple access—a novel non-orthogonal multiple access for fifth-generation radio networks[J]. IEEE Transactions on Vehicular Technology, 2017, 66(4): 3185-3196.

[9] WANG C L, CHEN J Y, CHEN Y J. Power allocation for a downlink non-orthogonal multiple access system[J]. IEEE Wireless Communications Letters, 2016, 5(5): 532-535.

[10] DING Z G, FAN P Z, POOR H V. Impact of user pairing on 5G non-orthogonal multiple-access downlink transmissions[J]. IEEE Transactions on Vehicular Technology, 2016, 65(8): 6010-6023.

[11] CHOI J. On the power allocation for MIMO-NOMA systems with layered transmissions[J]. IEEE Transactions on Wireless Communications, 2016, 15(5): 3226-3237.

[12] SUN Q, HAN S, I C L, et al. On the ergodic capacity of MIMO NOMA systems[J]. IEEE Wireless Communications Letters, 2015, 4(4): 405-408.

[13] LIU Y, ELKASHLAN M, DING Z G, et al. Fairness of user clustering in MIMO non-orthogonal multiple access systems[J]. IEEE Communications Letters, 2016, 20(7): 1465-1468.

[14] CHOI J. Power allocation for max-sum rate and max-min rate proportional fairness in NOMA[J]. IEEE Communications Letters, 2016, 20(10): 2055-2058.

[15] CHINGOSKA H, HADZI-VELKOV Z, NIKOLOSKA I, et al. Resource allocation in wireless powered communication networks with non-orthogonal multiple access[J]. IEEE Wireless Communications Letters, 2016, 5(6): 684-687.

[16] SUN Q, HAN S, I C L, et al. Energy efficiency optimization for fading MIMO non-orthogonal multiple access systems[C]// IEEE International Conference on Communications (ICC), June 8-12, 2015, London, UK. Piscataway: IEEE Press, 2015: 2668-2673.

[17] ZHANG Y, YANG Q, ZHENG T X , et al. Energy efficiency optimization in cognitive radio inspired non-orthogonal multiple access[C]//IEEE 27th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Sept 4-7, 2016, Valencia, Spain. Piscataway: IEEE Press, 2016: 1-6.

[18] ZHANG Y, WANG H M, ZHENG T X, et al. Energy-efficient transmission design in non-orthogonal multiple access[J]. IEEE Transactions on Vehicular Technology, 2017, 66(3): 2852-2857.

[19] LIU F, PETROVA M. Proportional fair scheduling for downlink single-carrier NOMA systems[C]// 2017 IEEE Global Communications Conference(GLOBECOM 2017), Dec 4-8, 2017, Singapore. Piscataway: IEEE Press, 2017.

[20] LI C, ZHANG Q, LI Q, et al. Price-based power allocation for non-orthogonal multiple access systems[J]. IEEE Wireless Communications Letters, 2016, 5(6): 664-667.

[21] WANG Z Q, WEN C C, FAN Z F, et al. A novel price-based power allocation algorithm in non-orthogonal multiple access networks[J]. IEEE Wireless Communications Letters, 2017(99): 1.

[22] FAN Z F, WEN C C, WANG Z Q, et al. Price-based power allocation with rate proportional fairness constraint in downlink non-orthogonal multiple access systems[J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2017, 100(11): 2543-2546.

[23] AFOLABI R O, DADLANI, KIM K. Multicast scheduling and resource allocation algorithms for OFDMA-based systems: a survey[J]. IEEE Communications Surveys & Tutorials, 2013, 15(1): 240-254.

[24] NG D W K, LO E S, SCHOBER R. Dynamic resource allocation in MIMO-OFDMA systems with full-duplex and hybrid relaying[J]. IEEE Transactions on Communications, 2012, 60(5): 1291-1304.

[25] VENTURINO L, ZAPPONE A, RISI C, et al. Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination[J]. IEEE Transactions on Wireless Communications, 2015, 14 (1): 1-14.

[26] DI B Y, SONG L Y, LI Y H. Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks[J]. IEEE Transactions on Wireless Communications, 2016, 15(11): 7686-7698.

[27] HOJEIJ M R, FARAH J, NOUR C, et al. Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access[C]// IEEE Vehicular Technology Conference (VTC Spring), May 11-14, 2015, Glasgow, UK. Piscataway: IEEE Press, 2015: 1-6.

[28] SUN Y, NG D W K, DING Z, et al. Optimal joint power and subcarrier allocation for MC-NOMA systems[C]// IEEE Global Communications Conference (GLOBECOM), Dec 4-8, 2016, Washington DC, USA. Piscataway: IEEE Press, 2016: 1-6.

[29] DI BY, BAYAT S, SONG L Y, et al. Radio resource allocation for downlink non-orthogonal multiple access (NOMA) networks using matching theory[C]//IEEE Global Communications Conference, Dec 6-10, 2015, San Diego, California, USA. Piscataway: IEEE Press, 2015: 1-6.

[30] LEI L, YUAN D, HO C K, et al. Joint optimization of power and channel allocation with non-orthogonal multiple access for

5G cellular system[C]//IEEE Global Communications Conference,Dec 6-10, 2015, San Diego, California, USA. Piscataway: IEEE Press, 2015: 1-6.

[31] LEI L, YUAN D, HO C K, et al. Power and channel allocation for non-orthogonal multiple access in 5G systems: tractability and computation[J]. IEEE Transactions on Wireless Communications, 2016,15(12): 8580-8594.

[32] SAITO Y, BENJEBBOUR A, KISHIYAMA Y, et al. System-level performance evaluation of downlink non-orthogonal multiple access (NOMA) [C]//IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Sept 8-11, 2013, London, UK. Piscataway: IEEE Press, 2013: 611-615.

[33] FU Y, SALALÜN L, SUNG C W, et al. Double iterative waterfilling for sum rate maximization in multicarrier NOMA systems[C]//IEEE International Conference on Communications (ICC), May 21-25, 2017, Paris, France. Piscataway: IEEE Press, 2017: 1-6.

[34] LI X, LI C, JIN Y. Dynamic resource allocation for transmit power minimization in OFDM-based NOMA systems[J]. IEEE Communications Letters, 2016, 20(12): 2558-2561.

[35] CAI W, CHEN C, BAI L, et al. Subcarrier and power allocation scheme for downlink OFDM-NOMA systems[J]. IET Signal Processing, 2017,11(1): 51-58.

[36] WEI Z, NG D W K, YUAN J. Power-efficient resource allocation for MC-NOMA with statistical channel state information[C]//IEEE Global Communications Conference (GLOBECOM), Dec 4-8, 2016, Washington DC, USA. Piscataway: IEEE Press, 2016: 1-7.

[37] SUN Y, NG D W K, DING Z, et al. Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems[J]. IEEE Transactions on Communications, 2017, 65(3): 1077-1091.

[38] FANG F, ZHANG H, CHENG J, et al. Energy-efficient resource allocation for downlink non-orthogonal multiple access network[J]. IEEE Transactions on Communications, 2016, 64(9): 3722-3732.

[39] FANG F, ZHANG H, CHENG J, et al. Energy efficiency of resource scheduling for non-orthogonal multiple access (NOMA) wireless network[C]//IEEE International Conference on Communications (ICC), May 22-27, 2016, Kuala Lumpur, Malaysia. Piscataway: IEEE Press, 2016: 1-5.

[40] TWEED D, PARSAEEFARD S, DERAKHSHANI M, et al. Dynamic resource allocation for MC-NOMA VWNs with imperfect SIC[C]// IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications, Oct 8-13, 2017, Montreal, Canada. Piscataway: IEEE Press, 2017: 1-5.

[41] ZHAI D, DU J. Spectrum efficient resource management for multi-carrier based NOMA networks: a graph-based method[J]. IEEE Wireless Communications Letters, 2017, 7(3): 388-391.

[42] KIM S H, CHAITANYA T V K, LE N T, et al. Rate maximization based power allocation and relay selection with IRI consideration for two-path AF relaying[J]. IEEE Transactions on Wireless Communications, 2015, 14(11): 6012-6027.

[43] KIM J B, LEE I H. Non-orthogonal multiple access in coordinated direct and relay transmission[J]. IEEE Communications Letters, 2015, 19(11): 2037-2040.

[44] MEN J, GE J, ZHANG C. Performance analysis for downlink relaying aided non-orthogonal multiple access networks with imperfect CSI over Nakagami-*m* fading[J]. IEEE Access, 2017(5): 998-1004.

[45] MEN J, GE J. Non-orthogonal multiple access for multiple antenna relaying networks[J]. IEEE Communications Letters, 2015, 19(10): 1686-1689.

[46] LIANG X, WU Y, NG D W K, et al. Outage performance for cooperative NOMA transmission with an AF relay[J]. IEEE Communications Letters, 2017, 21(11): 2428-2431.

[47] SUN H, WANG Q, HU R Q, et al. Outage probability study in a NOMA relay system[C]//IEEE Wireless Communications and Networking Conference(WCNC), March 19-22, 2017, San Francisco, USA. Piscataway: IEEE Press, 2017: 1-6.

[48] XU P, YANG Z, DING Z, et al. Optimal relay selection schemes for cooperative NOMA[J]. IEEE Transactions on Vehicular Technology, 2018, 67(8): 7851-7855.

[49] YUE X, LIU Y, KANG S, et al. Spatially random relay selection for full/half-duplex cooperative NOMA networks[J]. IEEE Transactions on Communications, 2018(99): 1.

[50] KIM J B, LEE I H. Capacity analysis of cooperative relaying systems using non-orthogonal multiple access[J]. IEEE Communications Letters, 2015, 19(11): 1949-1952.

[51] XU M, JI F, WEN M, et al. Novel receiver design for the cooperative relaying system with non-orthogonal multiple access[J]. IEEE Communications Letters, 2016, 20(8): 1679-1682.

[52] LIU X,WANG X, LIU Y. Power allocation and performance analysis of the collaborative NOMA assisted relaying systems in 5G[J]. China Communications, 2017, 14(1): 50-60.

[53] ZHANG S, DI B, SONG L, et al. Sub-channel and power allocation for non-orthogonal multiple access relay networks with amplify-and-forward protocol[J]. IEEE Transactions on Wireless Communications, 2017, 16(4): 2249-2261.

[54] NGUYEN T M, AJIB W, ASSI C. A novel cooperative non-orthogonal multiple access (NOMA) in wireless backhaul

two-tier HetNets[J]. IEEE Transactions on Wireless Communications, 2018, 17(7): 4873-4887.

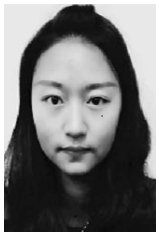
[55] DING F, WANG H, ZHANG S, et al. Impact of residual hardware impairments on non-orthogonal multiple access based amplify-and-forward relaying networks[J]. IEEE Access, 2018(99): 1.

[56] SELIM B, MUHAIDAT S, SOFOTASIOS P C, et al. Performance analysis of non-orthogonal multiple access under I/Q imbalance[J]. IEEE Access, 2018(6): 18453-18468.

[57] DENG C, ZHAO X, ZHANG D, et al. Performance analysis of NOMA-based relaying networks with transceiver hardware impairments[J]. Ksii Transactions on Internet & Information Systems, 2017(12): 134.

[作者简介]

王正强（1983−），男，博士，重庆邮电大学通信与信息工程学院副教授，主要研究方向为 5G 移动通信理论与关键技术、绿色通信、无线资源管理与优化。



成藻（1993−），女，重庆邮电大学通信与信息工程学院硕士生，主要研究方向为 NOMA 通信系统能效优化。

樊自甫（1977−），男，重庆邮电大学经济管理学院教授，主要研究方向为电信组织与运管管理、下一代网络技术。

万晓榆（1963−），男，博士，重庆邮电大学经济管理学院教授，主要研究方向为下一网络技术、通信运营管理。

基于平台战略的元器件分销生态圈建设

宋健

（中国中电国际信息服务有限公司，广东 深圳 518000）

摘　要：利用平台化思维，构建元器件分销的产业生态圈，通过对上游 IC 原厂及众多合作伙伴资源的有效整合，满足电子元器件产业下游广大的设计生产企业的长尾需求，是国内电子元器件厂商发展壮大的必由之路。从国内元器件分销的行业现状分析入手，通过国内某大型元器件分销商的平台转型实践，阐述了元器件分销商利用平台战略构建生态圈的重要价值和意义。

关键词：元器件分销；平台；生态圈；电商；转型

中图分类号：X32

文献标识码：A

doi: 10.11959/j.issn.1000−0801.2018226

Construction of ecosystem for electronic component distribution based on platform strategy

SONG Jian

China Electronics Information Service Co., Ltd., Shenzhen 518000, China

Abstract: Inspired by the platformized thinking, it is helpful for domestic electronic component distribution to build an industrial ecosystem by effectively integrating the upstream IC manufacturer and numerous partner resources to meet the long tail requirements of the downstream design and manufacturing companies, which can help the distributors to grow stronger and influential. Starting from the current situation analysis of domestic electronic component distribution industry, the necessity and importance of constructing ecosystem utilizing the platform strategy were explained through the practice of one large electronic component distributor in China.

Key words: electronic component distribution, platform, industrial ecosystem, e-commerce, transformation

1　中国元器件分销行业现状

1.1　元器件市场规模和增长速度

电子元器件处于电子信息产业链上游，是通信、计算机及网络、消费电子等系统和终端产品发展的基础，对电子信息产业的发展起着至关重要的作用。近年来，随着物联网、智能家居、AR/VR、可穿戴设备、汽车电子等新产业及新业态的发展，市场对电子元器件的需求不断增加，根据海关及相关机构数据统计，2013 年，中国电子元器件市场的采购规模已经超过 2 万亿元，2017 年超过 4 万亿元，2013—2017 年的复合增长率达到 19%，中国已经成为全球最大的电子元器件采购市场。

收稿日期：2018−05−03；修回日期：2018−08−02

1.2 元器件产业链上下游特征

目前我国从事电子制造的有大约 3 000 家大型企业和 300 万家中小企业，具有多样化的 IC 产品采购需求，采购份额相对分散，对 IC 原厂而言，其难以建立大规模的工程技术服务团队服务于数量庞大的客户；对电子产品制造商而言，其难以从相对集中的设计制造公司获得足够的应用技术支持，从而产生应用技术的市场缺口。从这个角度来看，电子元器件整体产业链呈现一个巨型的金字塔结构（如图 1 所示），上游 IC 原厂难以匹配下游广大生产企业的多样化需求，分销商作为链接上下游资源和需求的中游组织，在产业链中将发挥越来越大的作用，在亚洲，电子产品制造商通过分销商采购的份额为 51%~55%；而在北美、欧洲成熟市场，这个比例已达到 60%~80%。

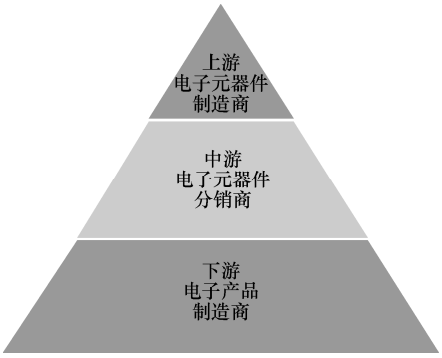


图 1 电子元器件产业链条

1.3 分销企业的国内竞争态势和发展方向

在我国，70%的元器件产品通过分销商采购。近年来，随着半导体价格的下降和生产成本的上升，元器件利润空间不断缩水，市场竞争越来越激烈，国内元器件的分销商集中度在提升，中等规模分销商将被整合或被迫缩减规模以提高竞争力，小规模分销商将靠小额业务找到他们的生存之道，未来将呈现大规模分销商和小规模分销商占主体的哑铃型竞争格局。

从分销商所面对的下游行业应用来看，物联网时代来临所导致的“长尾”效应逐渐显现，由于构建物联网的产品具有细分化、碎片化的特点，其对上游元器件的需求也更加趋向分散，分销商不能再像以前那样只重视市场上 2:8 分化中的大宗商品，众多小市场的汇聚，已经可以产生与主流相匹敌的市场力量，特别是近年来出现的创客群体更进一步推动了这种趋势的形成。目前，各分销企业越来越关注长尾行业中小型 OEM（原始设备生产商）企业的需求，期望通过平台化、线上化等手段建设在线交易平台，支持中小型企业开发适用于众多细分市场的产品。有专家预估，未来10年中国电子元器件市场中在线交易将占到整个分销市场规模的 10%左右。因此，针对长尾需求，如何利用平台化思维、信息化手段实现商业模式的创新，提升潜在的销售机会和行业影响，是元器件分销企业在新时代下转型升级需要主要考虑的问题。

2 平台战略对元器件分销转型的战略分析

2.1 平台战略的价值

平台是指由平台主提供基础资源及支撑服务构建的、多主体共享且具备正向网络效应的一种多方共赢的商业生态系统。平台模式的特点是至少连接两方或以上市场群体，通过单个群体内部及多方群体之间的相互依赖关系，构建无限增值的可能性，即产生正向的网络效应，如图 2 所示。传统的经济现象将消费时所获得的价值观视为个人层面的东西，与他人无关；然而在现实中却存在这样一些产品或服务，当使用者越来越多时，每一位用户所得的消费价值都会呈跳跃式增加，以微信为例，当全世界只有一个微信用户时，这个用户得不到任何使用价值，而当微信用户越来越多时，每一个用户可以通过微信联系的人数变多，其通过微信得到的使用价值会越来越多，这就是所谓的“网络效应”，微信用户所得到的使用价值随着用户数量的增加而增加的现象称为正向的“同边网络效应”，而随着微信用户数的增加，微信平台会吸引更多的服务提供者聚集，如

更多的公众号资源、更多的小程序等；反过来，随着微信平台可提供的服务种类越来越多，会进一步推动更多的用户选择微信，这种现象称为正向的“跨边网络效应”。

图 2　平台的特征[1]

所谓生态，生物学中生态系统的概念是易变的，有时，人们用它来描写特定的生态共同体，有时描述地球生物圈，有时则描述处于两级之间彼此受益的关系系统。同样，商业概念的生态是，用它来识别和培育那些具备巨大利益潜力的、成套的、相互交织的关系，而不要考虑这些系统是大还是小，关键是它们体现了新奇的概念，有益于消费者，有益于更好地组织和领导企业[2]。生态构建更看重商业本身的逻辑，比如产业结构、行业属性、需求特征、竞争地位等，以用户需求为导向整合多种异质性要素，给客户提供一个价值系统而非单个产品，与客户保持相对紧密的联系而非仅在交易那一瞬间，尽可能多地在多业务之间共享资源（包括用户资源、数据资源等），追求资源效益最大化，建立并巩固在该行业领域的地位。

平台战略的关键在于构造出一个丰富的、多层次的、协同的、多角色价值融合的“生态圈”[3]，激发正向网络效应，平台连接的任意一方的成长都会带动另一方的成长。平台战略通过把多种业务价值链所共有的部分进行优化整合，从而成为这些业务必不可少或最佳选择的一部分，作为为多方群体提供服务的基础，达到降低搜寻成本、交易成本、提升用户体验的目的，并以此形成新的兼具稳定性和扩张性的业务战略。

2.2　平台战略对元器件分销厂商的价值分析

哈佛大学教授马克·扬西蒂表示“未来的竞争不再是个体公司之间的竞争，而是商业生态系统之间的对抗。但凡在事业上取得持续辉煌的企业和组织，绝不是靠一己之力去谋求自身的发展，而是平衡地利用关联组织的能量和价值组成一个新的竞争平台”。

根据《共赢》[4]一书中对企业在不同环境适宜采用不同战略的研究，从企业所面临的市场变化和创新程度及上下游关系的复杂程度两个维度，将企业的战略选择分为 4 个部分：在市场变动较小且上下游关系比较简单的情况下，企业适合采用商品（commodity）战略，即通过生产通用商品获得市场收益，如矿泉水企业；在市场变动较大但上下游关系比较简单的情况下，企业适合采用利基（Niche）战略，即专注在某一细分市场获得相关收益，这时企业往往依托某一平台来开展其细分业务，如众多的微信公众号运营者；在市场变动不大但上下游关系比较复杂的情况下，企业适合采用支配战略，即进行产业链的垂直整合，掌控更多资源，获得更多收益，如一些传统企业的垂直并购；在市场变动较大，且上下游关系比较复杂的情况下，企业适合采用基石（Keystone）战略，即通过平台构建生态，更有效地对抗外部环境的变化，并通过平台合作伙伴及自身资源的整合，攫取最大价值，如阿里巴巴、微信等平台玩家。基于商业环境的匹配战略如图 3 所示。

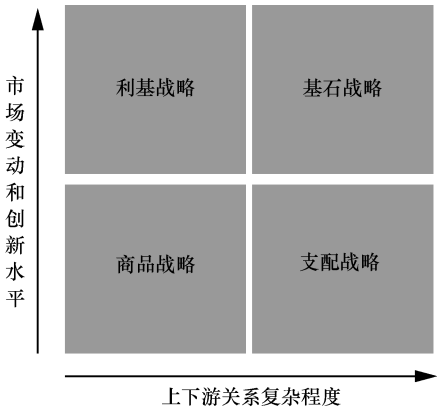


图 3　基于商业环境的匹配战略[4]

在物联网及人工智能等新技术层出不穷、创新创业团队不断涌现、市场需求越发个性和多元等综合环境背景下，电子元器件分销厂商面临一个复杂和创新的外部环境；同时，由于其位于电子元器件产业链中游，连接上游不同的 IC 原厂和下游不同的电子产品制造企业，下游成千上万家企业的个性化需求不断发生变化，单单依托上游的 IC 原厂资源很难有效及时匹配，需要元器件分销商综合产业中多方资源，满足下游的多元化需求，其所面临的上下游伙伴关系相对比较复杂；结合马克·扬西蒂的理论，元器件分销厂商适宜采用平台战略，并依托平台构建产业生态。

电子元器件分销商可通过构建元器件的在线交易平台（如图 4 所示），并依托线上服务的开展及交易数据的数字化，与物流、金融、科技等相关行业企业建立长效合作机制，逐步整合元器件交易、设计服务、社区服务、金融服务、云服务和大数据服务等在内的各项功能，逐步构建开放的电子元器件共赢生态圈，打造正向的同边网络效应和跨边网络效应，进一步吸引上下游市场群体向平台集聚，从而依托元器件分销构建产业生态，更有效地应对外部环境变化的影响，并提供多元化的业务组合，创造可持续性的经营收益，实现企业规模和企业价值的持续增长。

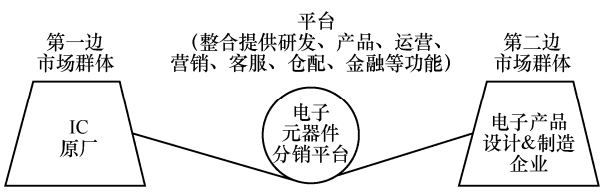


图 4　电子元器件分销平台

3　某大型元器件分销厂商 A 转型升级的实践分析

3.1　分销商 A 的平台发展策略

在目前的元器件电商平台中，主要以分销商自营模式为主，且尚未出现具有领导地位的企业，同时也缺乏集大成的生态圈平台。本文从卖方（自销、第三方卖家）构成、顾客触点、服务形式、盈利模式四大维度，对行业主要电商平台竞争对手——Avnet Express、开芯猫、科通芯城、Digi-key、ICkey 的能力进行比较，可以得出如下几个特点：第一，竞争者以分销商为主，现有电子元器件电商平台多数由传统分销商成立；第二，商业与运营模式单一，市场上没有一家集大成的平台，也尚未出现真正覆盖长尾市场的电商网站；第三，盈利模式单一，盈利模式以元器件与 PCB 销售为主，缺乏技术支持与供应链相关服务的增值收入；第四，竞争格局尚未形成，电子商务市场上缺乏占据领导地位的竞争者。

从市场环境和 A 的自身情况来看，笔者认为在当前阶段 A 建立开放式平台的业务时机尚未成熟，主要原因包括如下 3 个方面：首先，建立开放平台的市场影响力较弱，A 目前缺乏同时吸引供需双方形成电商开放式平台价值的号召力，一方面是由于对原厂的议价能力较低，另一方面是由于电商品牌在市场上知名度较低；其次，开放式平台模式以提供服务为核心，建设所需的服务能力仍需时间和资源的投入，包括金融、物流服务能力，一站式服务能力，正品鉴别和保障能力的提升等；第三，目前，客户线上采购的习惯仍需培养，同时，原厂对于长尾客户的重视程度仍待时间发展。

结合以上分析，A 分销商的平台发展战略计划分为两个阶段进行实施，如图 5 所示。

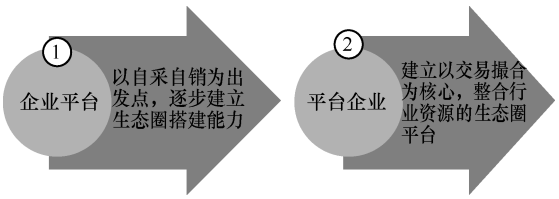


图 5　平台建设两步走战略

第一阶段是企业平台阶段。在这一阶段，A 重点构建企业自营的电商平台。通过以自采自销为出发点，逐步建立起生态圈的搭建能力，重视

电商核心能力的培育和发展，如数字营销能力、商品管理能力、物流能力、顾客体验管理能力等。

当第一阶段逐步成熟之后，可过渡至第二阶段——平台企业阶段。通过构建平台化的电商企业，建立以交易撮合为核心，整合行业资源的生态圈平台。其核心的发展需求包括构建合作伙伴的管理能力、数字营销能力以及其他类型的增值服务能力，包括供应链金融、技术服务、数据服务等。

3.2　分销商 A 的生态圈建设和实践

根据前述的平台发展策略，分销商 A 从 2013 年开始制定了面向平台化、数据化产业互联的转型升级路径，通过“能力构建”“资源整合”“生态共赢”逐步达成公司的战略规划，具体如图 6 所示。

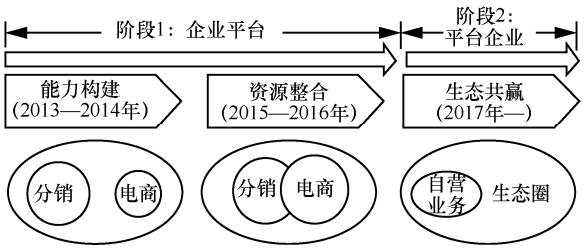


图 6　能力构建路径

（1）企业平台阶段

2013—2014 年，A 重点致力于能力构建，主要任务是由传统的分销模式逐步向电商模式过渡，在此阶段，打造电商平台，将分销业务逐步向电商平台转移，分销和电商同步发展：分销业务实现了扩大代理线、强化专业技术能力、完善客户布局、增加流量、提升品牌知名度、积聚稳定客源、引进新代理线等目标，而电商业务在此阶段实现了从初级到壮大的过程，流量不断增加，电商品牌知名度不断提升，通过借助内部代理资源和外部合作伙伴资源，丰富了电商合作产品线，提升了客户 BOM（bill of material，物料清单）满足率，实现了客源的稳定积聚和客户渗透率的持续提升。

2015—2016 年，经过能力构建，基本实现了达标的流量、稳定的中小型客户资源和相对丰富的产品线资源，在此基础上向“资源整合”过渡：将分销业务与电商业务进一步融合，并依托积累的客户资源向 IC 原厂争取更有竞争力的代理线，扩充产品线的授权销售区域、客户群，服务于大中小客户端到端需求以及进一步开发新产品线，提升销售额。在资源整合环节，创建致力于为中小型企业提供智慧设计链服务的部门，通过教育、样片、展会、微信等多种新技术手段发展会员，扩大覆盖，目前发展中小客户和会员粉丝近 20 余万，成为国内有相当知名度的双创服务品牌。并在此基础上，进一步拓展产品线代理权开发，国内外代理产品品牌超过 110 余条，在国内本土分销商中名列前茅。同时，在广东省东莞市虎门镇，开始了国内最大的元器件单体专业仓库的基建工作，为供应链的平台化转型奠定了物质基础。

（2）平台企业阶段

经过阶段 1 的实施，分销商 A 积聚了广泛的客户规模，拥有了丰富的产品代理线，具备了完备的物流、技术能力，实现了稳健的内部管理和人才储备。从 2017 年开始，A 开始了向“生态共赢”的方向转变。该阶段主要目的是构建供应链生态圈，通过线上线下一体化业务，进一步构建电商生态圈，提供多元化服务组合，实现企业规模和社会价值的持续扩大。2017 年，建成虎门大型元器件专业仓库，成立为元器件产业链提供智慧供应链服务的组织机构，对全行业提供开放的仓储和物流平台服务，汇集元器件交易大数据，推进产融结合，并在此基础上发展新型智慧供应链业务模式。目前，各项工作正在根据规划的思路稳步推进，以最终实现以元器件分销为基础，提供多种增值服务和合作模式的生态圈共赢体系。

4　结束语

A 作为元器件分销的专业公司，依托业务基础，用互联网的技术手段重塑业务流程，实现了业

务的平台化、数据化，很好地抓住了国内半导体产业的历史机遇，实现了快速增长。换句话说，通过对产业的信息化深度改造，实现了业务数据的收集、整理、分析、决策和上下游的数据价值增值。而从更宏观的层面来看，元器件分销的数据应该能够成为社会公共资源，进而对各行各业，比如银行、保险、运输、教育等形成数据的溢出价值，后续将通过持续构建合作，扩大生态圈范围，进一步挖掘元器件分销数据的使用价值和社会意义。

参考文献：

[1] 陈威如, 余卓轩.平台战略: 正在席卷全球的商业模式革命[M]. 北京: 中信出版社, 2013.
CHEN W R, YU Z X. Platform strategy: business model in revolution[M]. Beijing: CITIC Press Group, 2013.

[2] MOORE F J. 竞争的衰亡: 商业生态系统时代的领导与战略[M]. 梁骏, 译. 北京: 北京出版社, 1999.
MOORE F J. The decline of competition: leadership and strategy in the era of business ecosystems[M]. Translated by LIANG J. Beijing: Beijing Publishing House, 1999.

[3] 陈小玲. 开放平台生态体系及其第三方服务市场发展研究[J]. 电信科学, 2014, 30(8): 84-88.
CHEN X L. Ecological system of open platform and development of the third party service market[J]. Telecommunications Science, 2014, 30(8): 84-88.

[4] IANSITI M, LEVIEN R. 共赢[M]. 王凤彬, 等译. 北京: 商务印书馆, 2006.
IANSITI M, LEVIEN R. Win-win[M]. Translated by WANG F B, et al. Beijing: The Commercial Press, 2006.

[作者简介]

宋健（1968−），男，中国中电国际信息服务有限公司董事长、高级工程师，主要研究方向为企业战略、商业生态体系、供应链金融等。

基于 NFV 的边缘计算承载思路

罗雨佳，欧亮，唐宏
（中国电信股份有限公司广州研究院，广东 广州 510630）

摘　要：针对运营商如何引入边缘计算的问题，提出了一种基于 NFV 的边缘计算承载方案和部署思路。首先介绍了边缘计算、物联网等新兴技术产业的发展现状和趋势，梳理了边缘计算的概念、应用场景及具体需求，并阐述了 NFV 与边缘计算的关系；然后讨论了使用 NFV 对边缘计算平台进行承载和初步部署的思路，为后期边缘计算的规划和部署提供了技术参考。

关键词：通信技术；边缘计算承载方案；NFV
中图分类号：TN915
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018214

Bearing thinking of edge computing based on NFV

LUO Yujia, OU Liang, TANG Hong
Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: Aiming at introducing edge computing into operator’s network, a bearing and deployment thinking based on NFV was presented. Firstly, the industry development trends of edge computing were introduced. Then the concept, service scenarios and special requirements of edge computing were described, as well as the relationship between edge computing and NFV. Then, a bearing thinking of edge computing based on NFV was provided. Further more, some preliminary discussions about deploying edge computing platform on operator’s network were offered. Technical reference for the future deployment of edge computing was provided.

Key words: communication technology, bearing solution of edge computing, network function virtualization

1　引言

近年来，5G、物联网和边缘计算成为业界炙手可热的议题，这 3 个看似独立的方向有着深入的联系：边缘计算是 5G 的关键特征，可为物联网提供实时计算、分析和管理能力；而 5G 在架构、接口、信令和管理等方面，针对边缘计算和物联网需求做出了针对性改进，能更好地配合边缘计算与物联网业务。

随着传感器、智能家居、智能手机等智能设备的爆发式增长，未来将会有更多场景使用边缘计算。IDC 和 ITU-T 统计数据显示，到 2020 年，全球将部署近 2 120 亿个传感器，有超过 500 亿台终端和设备联网，每人每秒产生 1.7 MB 的数据量，其中超过 50%的数据需要在网络边缘侧存储、处理与分析；而到 2025 年，连接数将达到 1 000 亿，

收稿日期：2018−02−26；修回日期：2018−07−06

边缘计算市场呈现井喷式发展态势。

从2017年开始，各大巨头纷纷发力边缘计算，一些机构也参与到边缘计算的研究中。2016年11月，华为技术有限公司、中国科学院沈阳自动化研究所、中国信息通信研究院、英特尔、ARM和软通动力等多家公司联合成立了边缘计算联盟。2017年，亚马逊推出了应用于边缘计算的“AWS Greengrass”平台，微软也在其开发者大会上推出Azure IoT Edge，将云平台扩展到物联网边缘设备。2018年3月，Linux基金会发布了Akraino项目，旨在为运营商和企业网络构建边缘计算基础架构开发堆栈。

顺应产业趋势，运营商正在利用SDN/NFV、容器、微服务、CICD（continuous integration continuous deployment，持续集成持续交付）等新技术进行深层次的网络架构变革。边缘计算将是新的着力点，如何引入边缘计算，实现对边缘计算的承载，是运营商的重点研究课题之一。

本文首先介绍对边缘计算的理解，描述边缘计算的需求及与NFV的关系；然后给出NFV架构各层对边缘计算的承载思路，并探讨了边缘计算的部署问题，为后续边缘计算的落地提供技术参考。

2 边缘计算

2.1 边缘计算的概念

2015年，ETSI针对边缘计算方向成立了MEC工作组。初期，MEC的英文全称是mobile edge computing，强调移动性和无线网络环境，目前MEC已更名为multi-access edge computing，对多种接入方式和网络承载方式提供了支持。ETSI对边缘计算的定义是：在网络边缘提供IT应用和云计算能力，并保证近距离、低时延和高带宽。产业界对网络边缘的定义尚无统一标准，应根据各自需求和网络情况具体确定。对运营商来说，网络边缘主要指端局和接入机房。

边缘计算将计算、网络、存储能力下沉到靠近数据源头的网络边缘侧，构建了一种服务平台，就近提供边缘智能服务，旨在进一步减小时延，提高网络运营效率，提高业务分发/传送能力，优化终端用户体验。同时，部署于边缘计算平台上的各种业务，可利用从终端获取的网络或用户信息，提供更加个性化的服务。

论其本质，边缘计算实际是云计算的延伸。随着全球数字化浪潮的来袭，网络边缘到云数据中心的带宽和时延限制了传统云计算的表现；同时，云计算已经无法匹配来自各行各业海量数据的处理需求。因此，边缘计算和传统云计算必须相互协同，才能实现运营商业务的数字化转型。边缘计算作为前台，更靠近物理设备，可实现数据的快速采集和预处理；云计算作为后台，对非实时性数据进行价值分析，形成策略，为业务决策提供支撑。

目前，边缘计算主要有以下几大应用场景。

（1）智能视频加速

提升移动网络和固网用户对视频的访问速度，缓解快速增长的视频业务对现网造成的压力。

（2）密集计算辅

在网络边缘对云端计算提供辅助，减轻云数据中心压力，降低传输成本，提升性能。

（3）增强现实

配合增强现实（augmented reality，AR）摄像头数据和位置信息，对提升用户体验所需的额外信息进行更新，有效保障AR对实时性和数据处理精度的需求。

（4）物联网网关

提供低时延的流量分发、数据处理能力。

（5）车联网

更好更快地支撑车辆感知、娱乐、路况分析等车内应用。

（6）视频流分析

在本地对监控摄像头拍摄的数据进行分析。

（7）智能家居、智能制造

提升生产和控制效率，针对性地保障数据安全。

以上场景具备高可靠、低时延、高速度、本地化等特征，对边缘计算平台的具体要求如下：

- 灵活的基础设施承载；
- 良好的扩展能力；
- 敏捷连接，提供实时业务；
- 安全与隐私保护；
- 低能耗；
- 高水准服务质量和服务体验。

2.2　边缘计算的架构

ETSI 认为，边缘计算实际是一个开放的计算与感知控制服务平台，可部署多种应用。它既提供网络感知、计算、数据分析等服务，也为自营或第三方边缘应用提供虚拟化管理能力。

ETSI2016 年发布的 GS MEC 003 规范给出了边缘计算平台的参考架构，如图 1 所示，架构总体分为 3 层。

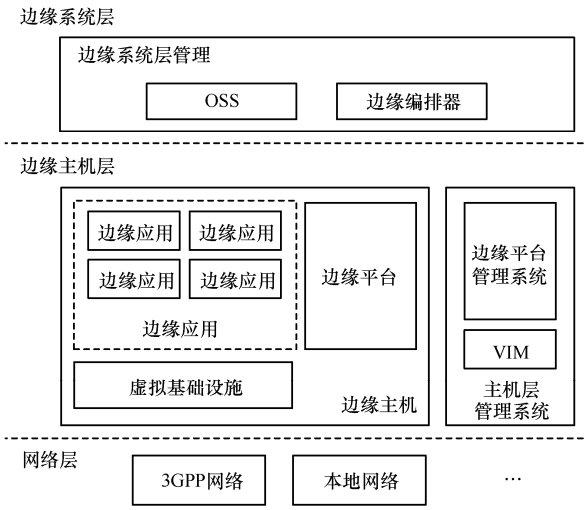


图 1　边缘计算参考架构

网络层用于提供进出边缘计算平台的管道，边缘计算支持多种接入方式，该管道可基于移动网，也可基于固网。

边缘主机层主要由边缘主机和主机层管理系统构成。其中，边缘主机基于虚拟化软件实现了资源池化，提供一套虚拟化的基础设施，可承载各种 5G、物联网相关的边缘应用软件，应用软件接受边缘平台的管理；而主机层管理系统包括边缘平台管理系统和 VIM（virtualized infrastructure manager，虚拟资源管理）系统，分别提供对边缘平台和虚拟网络、计算、存储资源的整体管理。

边缘系统层主要由边缘编排器和 OSS 组成，该管理系统主要提供全局的业务编排能力以及运营支撑能力。

2.3　边缘计算与 NFV 的关系

边缘计算与 NFV 的关系密不可分，ETSI 提供的边缘计算参考架构实际是参照NFV架构进行设计的。ETSI 认为，边缘计算可视为部署在网络边缘的本地业务网，类似于运营商的政企应用，其对资源的共享性及扩展性要求较高，需使用虚拟化环境进行部署。同时，边缘计算与 NFV 的本质，都是将各种应用软件运行在虚拟化平台之上，两者的底层基础设施乃至架构都是十分相似的，建议尽量复用 NFV 的环境和管理方案。

边缘计算并非是一个全新的、需要从头开垦的领域，对于运营商来说，应站在保护运营商现有投资、最大化利用现有经验、获取最大化收益的角度，基于已发展多年的 NFV 研究、开发、集成测试、现网实验、部署等经验，使用 NFV 环境实现对边缘计算业务的承载。

3　基于 NFV 的承载思路

运营商普遍采用的 NFV 架构包含 NFV 基础设施层、网络功能层和业务编排层，其架构如图 2 所示。

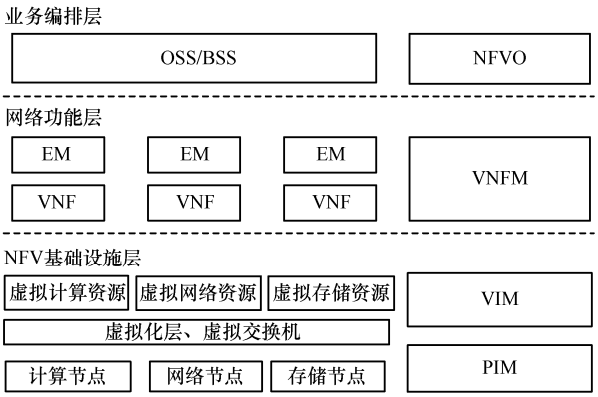


图 2　NFV 架构

使用 NFV 环境进行边缘计算承载，需重点讨论两方面内容：

- 应明确NFV架构功能模块与边缘计算参考架构模块间的对应关系，理清基于 NFV 的承载思路；
- 由于边缘计算平台具备第 2.1 节所述特点，因此其对 NFV 环境的硬件、软件存在某些特殊要求，应做出针对性调整和优化。

接下来，本文将讨论 NFV 不同层面对边缘计算的承载方案。

3.1 NFV 基础设施层

NFV 基础设施层主要包括计算节点（通用服务器）、网络节点（交换机）、存储节点（硬盘、磁阵等）、虚拟化层、虚拟资源及两大管理系统 VIM 和 PIM（physical infrastructure manager，物理资源管理）系统。

在 ETSI NFV 工作组早期定义中，NFV 基础设施不包含 VIM，然而目前业界普遍认为 VIM 也属于 NFV 基础设施的范畴，因为 VIM 需要与物理和虚拟资源进行密切交互。在物理资源的管理方面，虽然现有 VIM 产品包含一部分物理硬件管理功能，但功能尚不完善，无法满足运营商对不同厂商设备统一管控的需求，因此需要使用独立的 PIM 系统，实现对服务器、交换机和存储设备的统一管理。

NFV 基础设施的计算、网络、存储节点、虚拟化层（含虚拟交换机）对应边缘计算参考架构的虚拟基础设施，边缘计算架构中的 VIM 功能与 NFV 相同。同 NFV 的思路类似，运营商应当在边缘计算架构中增加 PIM 系统，作为主机层管理系统的独立子功能。

运营商网络边缘不同机房的条件存在较大差距，大部分机房在面积、供电、制冷、承重方面具有一定限制，考虑到边缘计算业务的特殊要求，NFV 基础设施层需在以下两方面进行改进和优化。

（1）硬件设备选型

建议在保证可靠性的前提下，配备所需物理硬件的最小集合，并选择低功耗、占用空间小、重量轻的设备。

（2）VIM、PIM 优化部署

NFV 化后，由于 OpenStack、分布式存储等系统存在可靠性要求，需使用多台服务器进行冗余部署，实际引入了更多的物理设备，部分对空间和功耗非常敏感的边缘机房无法支持更多设备的引入，应考虑将 NFV 管理系统 VIM 和 PIM 进行优化部署，降低其对机房资源的消耗。

3.1.1 硬件设备的选型

边缘计算硬件设备的选型应本着最大化利用现有端局和接入机房环境、减少部署成本、降低 NFV 化改造施工复杂度的原则，对服务器尺寸、类型进行限制，并结合业务需求，给出不同类型的服务器典型配置，详见表 1。

表 1 的具体数值要求，参考了行业标准 YD/T 1821-2008[1]、企业标准《中国电信 IDC 机房设计规范》[2]和《中国电信单机定制化服务器工程总体技术要求》，考虑到端局和接入机房的环境限制，并结合业界服务器产品的实际情况，本文对服务器尺寸和配置进行了调整，该配置已在中国电信网络重构机房的各种测试中得到了验证，具备一定的普适性。

总体来说，可将服务器分为以下 3 种类型。

- 计算型：适用于数据分析计算业务。
- 转发型：适用于对时延、吞吐量要求较高的实时性大流量业务。
- 存储型：适用于数据存储业务。

上述 3 种模型是服务器的基本分类，根据边缘计算业务的实际需求，服务器的配置可能是多种类型的组合：比如智能视频加速对磁盘容量和 IOPS 都有较高要求，对应的服务器配置应该是转发型和存储型的集合。

表 1　服务器选型建议

参数	接入机房	端局
服务器类型	建议采用机架服务器	
服务器宽度	建议宽度不超过 500 mm	
服务器深度	建议深度不超过 700 mm	建议深度不超过 950 mm
计算型服务器	可采用 1U 或 2U 服务器； CPU 2×12 核以上，主频不低于 2.2 GHz； 内存 256 GB 以上； 建议至少插入 2 块支持 DPDK 的 PCIe 万兆网卡	
转发型服务器	应采用 2U 服务器，提升 PCIe 卡槽可扩展性； CPU 2×10 核以上，主频不低于 2.2 GHz； 内存 256 GB 以上； 可插入 3 张支持 DPDK 的 PCIe 网卡，端口逐步向 40GE/100GE 演进； 功耗不超过 400 W； 建议服务器提升对高温的耐受能力	应采用 2U 服务器，提升 PCIe 卡槽可扩展性； CPU 2×10 核以上，主频不低于 2.2 GHz； 内存 256 GB 以上； 可插入 6 张支持 DPDK 的 PCIe 网卡，端口逐步向 40GE/100GE 演进
存储型服务器	CPU 2×10 核以上，主频不低于 2.2 GHz； 内存 128 GB 以上； 考虑成本、体积和重量因素，可选用低功耗、重量轻的 SSD 硬盘实现分布式存储	CPU 2×10 核以上，主频不低于 2.2 GHz； 内存 128 GB 以上； 可选分布式存储或磁阵，根据数据重要性、IOPS 要求按需选择

3.1.2　VIM 和 PIM 的优化部署

对于规模较小、基础设施条件受限的网络边缘机房，若每个机房都要部署 VIM 和 PIM 管理系统，会造成巨大的资源浪费，建议采用本地精简部署方式或远程部署方式。

本地精简部署方式通过将计算节点和 VIM/PIM 统一部署，并采用裁剪 VIM/PIM 部分组件等技术手段，节省 VIM/PIM 所占用的资源。该方案对网络规划要求较高，需做好 VLAN、VxLAN 划分，有效隔离计算节点业务流量，保证边缘计算业务不受统一部署影响；同时，应合理分配服务器计算、存储资源，防止管理系统对资源的过多占用，导致用户体验的降低。

远程部署是将 VIM 和 PIM 部署在机房条件限制较少的机房，对多个网络边缘机房的物理资源、虚拟资源进行统一管理。该部署方案受到消息通道的限制，开源 OpenStack 消息总线的大小限制了 VIM 可以管控的服务器数量，并且 OpenStack 内设消息计时器，若超过一定时间未收到服务器应答，则判定连接故障，继而引发系统重启操作，所以对 VIM 的远程部署距离存在限制。如果要实现该方案，需联合厂商对 VIM 的进行相应调整，并在现网实际部署测试，以验证其可靠性。

3.2　NFV 网络功能层

VNF（virtualized network function，虚拟网络功能）可承载如 vBRAS、vIMS 等运营商虚拟网元，也可承载边缘计算应用，两者都是安装在虚拟环境之上的应用软件。

在 NFV 中，EM（element manager，网元管理）系统提供对虚拟网元的业务和资源管理，可对应于边缘计算架构中的边缘平台。不过在 ETSI 的要求中，边缘平台不仅需要管理应用软件的资源和业务，还需要执行流量策略控制，向数据平面下达命令。因此，若使用 EM 来承载边缘平台，需在原有基础上进行功能增强，以满足边缘计算对路径控制的需求。

VNFM（VNF manager，虚拟网络功能管理）系统提供对虚拟网元的生命周期管理。VNFM 可对应边缘计算架构中的边缘平台管理系统，在实际使用过程中，EM 和 VNFM 需要相互协同配合，共同完成对边缘计算应用的管理。

由于边缘计算应用具备高可靠、低时延等特征，要求边缘应用软件及 EM、VNFM 管理系统轻量化，具备快速响应、简易交互的能力；因此，建议使用支持云原生的边缘应用，并对现有 EM、VNFM 进行相应升级改造。云原生软件基于微服务实现，支持容器化部署。软件被分解为多个基础的原子功能，减少系统冗余，提升功能利用率。同时，任何错误和故障只会导致特定功能无法执行，不会对软件其他部分产生连带影响，并且便于故障定位，大大提升软件可用性和灵活性。

目前，云原生概念已在 NFV 产业中被广泛接受，各大厂商的 NFV 产品均进行了相应改进，相信到了边缘计算部署应用时，云原生方案会更加成熟。

3.3 NFV 业务编排层

NFV 的业务编排层与边缘计算架构中的边缘系统层可完全对应。由于该层面部署在运营商网络中相对较高的位置，主要从宏观角度对区域或全网进行管理和编排，因此受边缘计算业务特点和需求影响较小，无须做较大改动。

但是产业界中云原生概念已逐步发展并开始影响 OSS 和编排器。Linux 基金会旗下的开源项目 ONAP 正致力于实现智能、敏捷的网络管理和编排系统，该社区一直保持较高活跃度，它基于微服务架构，支持容器化部署，代表了网络编排和管理系统的发展趋势。

另外，从 2016 年开始，越来越多的厂商开始关注基于 AI 的网络编排管理，利用 AI 和大数据技术，帮助管理者进行数据分析和策略制定，从而实现更精确、更自动化的网络管理。2018 年年初，Linux 基金会联合 AT&T 成立了 AI 开源项目 Acumos，构建一个管理 AI 和机器学习应用程序，并共享 AI 模型的联合平台。它提供了可视化工作流程，支持自由共享 AI 解决方案和数据模型，这无疑将加速 AI 在网络编排管理系统中的应用。

未来，云原生和 AI 智能将辅助网络管理和编排系统，提升运营商网络管理能力，更灵活快捷地开通新业务，以适应市场和用户的新需求。

4 边缘计算的部署

对运营商来说，基于 NFV 的边缘计算平台可考虑部署在城域网端局或接入层，具体部署在哪个层面，需要紧密结合业务需求和现网实际情况。

对于智能视频加速，增强现实、物联网网关、智能家居等实时性要求非常高的业务，考虑尽可能靠近用户，将边缘计算平台部署于运营商接入机房，与固网 OLT 网元位于同一平面；同时，建议将 5G UPF 网关下沉到接入机房，减少业务传输时延。由于接入机房的空间、制冷效果、功率控制、承重等基础设施能力受限，可部署的设备规模较小，在不影响边缘计算平台管理实时性的前提下，可考虑将 VIM、PIM、VNFM 的部署位置适当提高到城域网边缘，不占用接入机房资源，实现集中式管控。目前这种方式只是一种实现思路，其合理性需要在现网实验中进行进一步验证。

对于本地数据分析、密集计算辅助等业务，时延要求相对较低，主要看重网络边缘的分析计算能力，因此可考虑将边缘计算平台部署在位置相对较高的城域网边缘。

边缘计算有一些独特的管理要求，如用户的移动会触发应用迁移、应用状态更新等，因此，初期建议采用独立的机房部署边缘计算平台，避免与运营商其他类型的业务混用机房，加重运维管理复杂度。为了最大化利用计算、网络和存储资源，建议将存储型边缘应用（如 CDN）和计算型边缘应用（如数据分析）部署在同一机房。后期，当边缘计算的标准和应用相对成熟后，可考

虑与运营商其他业务进行综合部署。

5　结束语

边缘计算、物联网、5G 等技术是近几年通信行业新的爆发点，各大运营商都在对相关技术、标准和产品进行研究和测试。随着运营商网络重构脚步[3]的不断加快以及对 NFV 等新技术理解的不断深入，如何使用新的网络架构实现对新型业务的承载，必然是亟待解决的问题。

本文提供了基于 NFV 的边缘计算承载思路，并探讨了边缘计算在现网中可能的部署方案。目前，边缘计算尚处于初期阶段，标准组织的规范制定刚刚起步，业界也缺乏可大面积推广的成熟应用，本文初步提供了一种边缘计算的承载和部署思路，还有待后续的深入研究和测试，逐步对解决方案进行完善。

参考文献：

[1] 工业和信息化部. 通信中心机房环境条件要求：YD/T 1821-2008[S]. 2008.
Ministry of Industry and Information Technology. Communication center equipment room environmental requirements: YD/T 1821-2008[S]. 2008.

[2] 中国电信集团公司. 中国电信 IDC 机房设计规范[S]. 2011.
China Telecom. China Telecom IDC engine room design specification[S]. 2011.

[3] 陈华南，龚霞，朱永庆，等.城域网重构思路[J]. 电信科学，2018, 34(5): 120-126.
CHEN H N, GONG X, ZHU Y Q, et al. Metropolitan area network re-architecture[J]. Telecommunications Science, 2018, 34(5): 120-126.

[作者简介]

罗雨佳（1989−），女，中国电信股份有限公司广州研究院工程师，主要从事 IP 承载网络技术、SDN/NFV 技术相关研究等工作。

欧亮（1968−），男，博士，中国电信股份有限公司广州研究院高级工程师，长期从事电信网络规划设计、互联网新技术研究与应用工作。

唐宏（1974−），男，中国电信股份有限公司广州研究院数据通信研究所所长，主要从事 IP 承载网、下一代互联网、网络新技术方面的研发与管理工作。

个人信息保护的利益衡量与制度构建

李美燕

（国家计算机网络应急技术处理协调中心，北京 100029）

摘　要：海量数据时刻记录着人们生产生活的轨迹，通过分析、挖掘这些数据可为网民提供个性化、精准性的服务。但数据利用的力度越大，个人信息保护面临的风险就越大，失衡现象就越发严重和常见。探索中国个人信息保护之道，需要符合数字中国的实践需要，需要结合中华民族伟大复兴的宏大背景进行具体制度的安排。首先以美国、欧盟、我国为例介绍个人信息保护的立法路径与模式，然后分析了我国个人信息保护的特点，最后对个人信息保护的利益平衡与制度进行了思考。

关键词：个人信息保护；人格权益；经济权益；网络安全；数字中国

中图分类号：D923

文献标识码：A

doi: 10.11959/j.issn.1000−0801.2018235

Interest measurement and system construction of personal information protection

LI Meiyan

National Internet Emergency Center, Beijing 100029, China

Abstract: Massive data always records the trajectory of people’s production and life. Through analysis and mining, it can provide personalized and accurate services for netizens. But the greater the intensity of data utilization, the greater the risk of personal information protection, and the imbalance is more serious and common. Exploring the way of protecting personal information in China requires meeting the practical needs of digital China. Firstly, taking United States, European Union, and China as examples, the legislative path and mode of personal information protection were introduced. Then, the characteristics of personal information protection in China were analyzed. Finally, the balance of interests and system of personal information protection was considered.

Key words: personal information protection, personality right, economic interest, network security, digital China

1　引言

当前，以互联网为代表的信息通信技术和人类生活生产交汇融合，互联网快速普及，消费互联网百花齐放，产业互联网跨界融合并持续激发创新浪潮，企业竞争力和影响力持续提升，互联网用户和市场规模日益庞大，海量数据的集聚对经济发展、社会治理、国家管理、人民生活都产生了重大影响。正是这些海量数据使人工智能得到广泛应用，使深度学习成为可能，企业则通过数据的分析、挖掘为网民提供个性化、精准性的服务。人们主动拥抱数字化生产生活的过程，也

收稿日期：2018−07−03；修回日期：2018−08−10

意味着对个人信息使用权的主动让渡。数据利用的力度越大，个人信息保护面临的风险就越大，失衡现象就越发严重和常见。个人信息的滥用、泄露引起了社会各界的广泛关注，也成为各国立法关注的热点。从立法层面来看，世界各国纷纷在个人信息保护方面做了前瞻性布局。欧盟地区从个人权利角度论证个人信息保护的必要性；美国则更关注个人信息的经济属性，将个人信息保护作为风险管理来对待；我国个人信息保护的框架体系雏形已经显现，但需要合理的制度设计和实现路径予以保障，其关键是要处理好 3 个平衡之间的关系。

2　个人信息保护的立法路径与模式

个人信息保护和数据治理成为全球的新问题，也成为各国立法关注的热点。截至 2017 年，全球已经有 120 个国家或地区先后颁布个人信息保护法律。2018 年，欧盟《一般数据保护规范》正式实施，美国《2018 加州消费者隐私法案》签署生效，我国个人信息保护立法尽快出台的呼声进一步高涨。从个人信息保护的权益关系来看，传统法律框架将个人信息纳入隐私权或通过一种独立人格来加以保护，并通过个人信息的自决功能等特殊效力，以用户知情同意的方式建立数据流通的关系，当然还有其他管理需要的例外情形。随着时代变迁和产业发展，个人信息更多地强调其流通、利用的价值，也就是其经济属性。如何适应这种形势，需要理清其中关系和平衡权益。

2.1　美国模式

美国模式以分散式立法为特点，即在各个行业分别制定有关个人信息保护的法律规则、准则，而不制定统一的个人信息保护法律。从立法层面来看，个人信息被置于隐私的范畴而加以保护，更重视对公共领域政府机关涉及利用个人数据的行为进行规范，如 1974 年的《隐私法》和 1980 年的《隐私保护法》[1]。同时，针对敏感信息特别保护需求进行特别立法，如 2013 年的《儿童在线隐私权保护法案》和美国加利福尼亚州的《商业和专业条例》。这种立法模式与美国法律隐私权概念的开放性有关，在实践中，美国从实用主义出发，并未对个人信息和隐私权进行严格分界[2]，更关注个人信息的经济特征和个人价值[3]，将个人信息保护作为风险管理来对待，采用行业自律和市场调节机制来实现。这种设置有利于信息的流通和利用，将个人信息的搜集、利用交由企业，由其与权利主体通过合同进行协商解决。但是，由于个人与企业所处地位、掌握信息的不对等，企业相继出现了不当收集、使用和移转个人信息的情况，从而使个人的权利难以获得全面充分的保护。在实践中问题就折射出来，如 Facebook 数据泄露事件，引发民众对个人信息保护的质疑和巨大的担忧。在此背景下，美国加利福尼亚州的《2018 加州消费者隐私法案》获得民众高票通过。该法案被称为美国“最严厉、最全面的个人隐私保护法案”，不仅大幅扩充适用范围，还创建访问权、删除权、知情权等一系列消费者隐私权利，进一步加重企业保护个人信息的责任。

2.2　欧盟模式

欧盟模式以制定统一法为特征，采取个人信息人格权保护模式，对数据处理者和控制者进行严格规范。在欧洲，多国曾尝试制定个人数据法，其中德国特征最为明显。德国联邦议会于 1977 年生效《联邦数据保护法》，第一次系统地、集中地保护个人信息，并彰显出其民事权利的属性。1981 年欧共体制定《关于自动化处理的个人信息保护公约》对大规模的自动化信息处理活动进行规范[4]。基于欧盟国家个人数据流动的实际情况，决定在一体化进程下统一个人数据保护立法。1995 年，欧盟通过《关于在个人数据处理过程中保护当事人及此类数据自由流通的 95/46/EC 指令》（简称《95 指令》），确立了个人信息保护的价

值，包括“基本权利”“自由”以及“隐私”的概念，构建了查阅权、更正和删除权、反对权、免受完全自动化决定权等权利。1997 年欧盟颁布《有关电信行业中的个人数据处理和隐私权保护的 97/66/EC 指令》（简称《97 指令》），适用于电信行业。2002 年《电子通信隐私指令》取代《97 指令》，要求电信和互联网服务商确保个人数据安全，确定了存储和使用数据时的主体同意规则。2009 年的《欧洲 Cookie 指令》对 cookie 使用和必要信息披露进行规范和管理。

但是上述规则面临着适用统一性的困境，并随着公众对数据安全质疑的进一步增强，欧盟开始着手制定更加一致、细致的数据保护框架，以提高个人数据控制能力，并规范数据利用市场。自 2010 年起，欧盟启动《个人数据保护指令》的修订计划，最终于 2016 年通过《一般数据保护条例》（GDPR）。GDPR 于 2018 年正式实施，适用范围大幅扩大，进一步明确数据主体的“知情同意”原则，细化并扩展了《95 指令》的查阅权、更正与删除权、反对权以及免受完全自动化决定权的内容，并增设限制处理权、可携带权、遗忘权[5]。

比较两种模式的立法经验，美国关注个人信息的经济特征，更多地将个人信息保护作为风险管理来对待，采用行业自律和市场调节机制为主的松散式立法。以欧盟为代表的国家（地区）从个人权利角度论证个人信息保护的必要性，尤其是 GDPR 对全球立法的推动与影响是巨大的，大有成为主导型的立法模式之发展态势。正如维克托教授所说，“当世界开始迈向大数据时代时，社会也将经历类似的地壳运动。在改变人类基本的生活与思维方式的同时，大数据早已在推动人类信息管理准则的重新定位。然而，不同于印刷革命，人没有几个世纪的时间去适应，人们也许就只有几年的时间”[6]。在这样的时代背景下，给我国个人信息保护立法留下的窗口期也不多了，民众的诉求和产业的需求持续高涨。从全球视野来看，不论是欧盟立法模式，还是美国立法模式，都有其合理的地方，都有各自的价值观和社会基础作为支撑。传统不能涵盖一切，它只能划一条模糊的界限，其内容不明确[7]。我国个人信息保护的立法路径的关键不在于模式差别，而在于制度设计是否符合中国文化的需要和产业发展的实际。

2.3 中国模式

我国个人信息保护立法之初走的是一种逆常规路线，即“最后法先行、刑法先行”。随着产业的蓬勃兴起，立法日益结合中国互联网产业高速发展的实践需要和中华民族伟大复兴的宏大背景。从刑法视角来看，顺应发展需要及时扩展个人信息的内涵，划出权利保护的红线、底线，并形成了刑法倒逼之势，促进行政立法、民事立法完善。从行政法视角来看，将个人信息作为产业运行安全、国家网络安全的重要组成部分。从民法视角来看，个人信息保护从重归属向重利用或者归属和利用并重的方向发展。具体来看，具有以下特点。

（1）及时回应个人信息和隐私保护的迫切需求

2012 年全国人民代表大会常务委员会《关于加强网络信息保护的决定》规定，涉及公民个人隐私的电子信息以及可识别的公民个人身份信息受国家保护，任何组织和个人不得非法获取、出售公民的个人信息。2015 年《刑法修正案（九）》将“出售、非法提供公民个人信息罪”和“非法获取供个人信息罪”整合为“侵犯公民个人信息罪”，放宽了侵犯公民个人信息罪主体范围。2017 年《关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》在《网络安全法》的基础上，进一步扩大了个人信息范围，将其特定自然人活动情况的各种信息（如行踪轨迹信息等）纳入保护的范畴。从立法进程来看，《中华人民共和国刑法》首先划出了一

条“高压线”。通过应该前置的低度的法先行，让秩序法优先发挥管理作用，使法律发挥规范、引领、保障的作用。然而，我国为何形成了“最后法先行”局面呢？这反映了一个很迫切的现实，如个人信息泄露助长电信诈骗、人肉搜索酿成生活悲剧。个人信息的滥用或对公民个人信息破坏、侵犯到了一个难以容忍的地步，迫使《中华人民共和国刑法》不得不出手。同时，从刑法视角来看，个人信息内涵也及时回应了产业发展需要，特别是人工智能应用服务纵深延展，从身份属性的信息向活动情况的各种信息方向延展。

（2）坚持信息和数据安全以及其他公共利益的底线

2013 年工业和信息化部《电信和互联网用户个人信息保护规定》要求电信业务经营者、互联网信息服务提供者应当对收集和使用的用户个人信息的安全负责。2017 年《中华人民共和国网络安全法》正式实施，将个人信息保护作为网络安全的重要组成部分，并明确了经营者在发生数据安全事件向用户和主管部门“双报告”的规定。保护个人信息是维护国家安全和公共安全的基础，我国也不例外，在立法中加重信息控制者和处理者在接触个人信息过程中的安全责任。

（3）兼顾商业利用与产业发展的需要

2017 年《中华人民共和国民法总则》第 111 条就对个人信息权益进行专门的确认和保障，第 127 条对数据和网络虚拟财产进行保护和调整。在我国民事基本法和单行法中首次出现了个人信息保护的有关规定。过去立法，权利概念中比较重归属。但是，随着我国产业发展和人工智能应用推进，《中华人民共和国民法总则》第 111 条与第 127 条，把个人信息保护和数据保护进行分置规定，既明确了个人信息的归属和保护问题，还考虑了数据开发利用和保护问题，将个人信息保护转向重利用或者归属、利用并重的方向延展。

总体来看，我国个人信息保护虽然没有统一立法，但已经形成了比较体系化的规定，表现出了先进性。当然，立法目前的先进性更多地表现在原则以及理念层面。

3　个人信息保护的利益平衡与制度思考

我国个人信息保护的制度设计超越了欧美两种模式的简单对比或移植，进一步回应个人信息和隐私保护的迫切需求，坚持信息和数据安全以及其他公共利益的底线，兼顾数据商业利用与产业发展的需要，其核心是相关主体的权利及权利体系的安排。在利益格局下，立法需要考虑个人信息在社会的作用并根据比例性原则与其他基本权利保持平衡，结合数字中国和中华民族伟大复兴的宏大背景来进行考虑。

（1）个人权益的保障与数据商业利用之间的平衡

用户的人身特征、行为状态时时被海量数据记录，尤其是指向特定主体的个人信息，能够轻松地勾勒用户的人格形象，显现其生活轨迹。用户发出的每一条信息、浏览的所有痕迹，都不再是自己独有的存储记录，不会被时间磨损，将清晰地成为善意或恶意的他人通向他的路标。个人信息的利用增进了社会福祉，也导致了信息主体权益受到威胁和侵害。个人信息保护的重要性被描绘得具体而真切。马斯洛在其需要层次理论中指出，“人格标识的完整性和真实性是主体受到他人尊重的基本条件”[8]。在具体制度构建中，个人信息对于权利主体的尊严和自由价值应当首先被考虑，即人格权益的保护，需要赋予用户信息知情权、信息决定权、信息更正权等权利。另一方面，个人信息保护需要兼顾数据商业利用的需要。毕竟，数据的流通使人们生产生活彰显便捷性，决策更加智能化，服务获取变得精准化。就用户而言，权利配置需要确立其人格权益，还要兼顾信息的经济属性，配置财产权益；就经营者

而言，分别配置数据经营权和数据资产权。因为没有经营者投入大量的人力、物力、财力，数据利用、挖掘以及人工智能应用难以延展，用户也无法时时获取免费的网络服务[9]。

同时，在具体制度的安排过程中要保持一种动态与弹性，以平衡个人权益保障和数据商业利用之间的关系。就人工智能应用而言，就会推翻很多过去确立的个人信息保护规则。在当下的产业环境下，企业收集的数据越多，分析能力越强，消费需求越大，商业利润就越多，以互联网公司通过 cookie 技术实现用户画像为例。虽然可以基于算法对人的工作表现、经济状况、兴趣爱好、行为习惯、位置等进行分析，并精准评价，但是在应用的过程中可能会出现性格歧视、隐私侵犯等问题。从个人信息保护的角度，赋予用户知情同意权利十分必要。但在具体的权利构建的过程中，仍需要处理好利益平衡的关系。比如，采用欧盟 GDPR 的“明示同意（opt-in）”方式，就难以满足企业收集海量用户数据的连贯性需求，在一定程度上会阻碍人工智能应用的创新发展。基于合理的数据商业利用目的，在非高风险数据或非敏感数据的应用场景下选择“退出同意（opt-out）方式”就能够较好地处理二者的关系。同时，知情同意原则还需要结合产业发展实际，设定例外原则。比如，新加坡立法，在互联网应用场景下所有数据的收集需要征得用户同意变得不现实的情况下，在符合合法商业目的的情况下，可以不征求用户同意，只要对例外情形进行安全影响评估即可。可以看出，日新月异的技术发展给个人信息保护带来直接的冲击和调整，迫使用户知情同意原则保持动态和弹性。

总之，我国个人信息保护既要及时构建权利保护屏障，也要使数据“物尽其用”，给产业创新与发展留有空间。只有保持数据的流通，才能以数据为纽带促进产学研深度融合，形成数据驱动型创新体系和发展模式，培育造就一批国际领军的互联网企业，筑牢数字中国之基。在个人信息保护的制度安排上，要坚持个人权益保障与数据商业利用并重，保持规则的动态与弹性，最终才能让中国互联网的发展成果造福更多国家和人民。

（2）公权力与私权利之间的平衡

在个人信息保护的具体制度构建中需要兼顾公权力与私权利之间的平衡。信息社会，不同于传统隐私保护中政府超然的中立地位，在个人信息保护和利用中，政府具有了利用者和管理者的双重角色：一方面，政府作为社会管理和社会福利的承担者，公共安全、公共管理和公共福利的推进离不开对居民个人信息的掌握；另一方面，出于对行政效率的追求，也不断促进政府积极探索个人信息利用的限度和价值[10]。

作为信息的利用者，政府不能无节制地、肆意地收集和利用个人信息。也就是说，公权力需要受到必要的限制。作为信息的管理者，在特定的情况下，需要对个人信息私权利进行必要干预。比如，为了公共安全，国家对跨境数据设定本地化存储的要求或要求跨境数据转移符合本国对个人信息保护标准。虽然互联互通的产业背景以及商业模式的创新迭代都需要数据的开放，但是基于公共安全对私权利进行必要限制，也是无可非议的。比如，为了公共利益或保护公众的知情权，不赋予相关主体删除有关数据权利或遗忘权。日本一名因猥亵儿童的罪犯起诉谷歌，依据遗忘权要求 Google 删除他被捕的信息，最后日本最高人民法院判决 Google 不需要删除上述信息。在本案中，相比删除权，保护公众的知情权就显得更为重要。再比如，欧盟 GDPR 第 112 条就规定为了完成《日内瓦公约》的义不容辞的任务或遵守适用于武装冲突的国际人道主体法，因为公共利益的重要原因或为了数据主体的切实利益，任何向国际人道主义组织客观上或法律上无法做出同意的数据主体的个人数据的传输可以被认定为是必

要的[11]。也就是说，为公共利益而传输、共享特定主体的个人信息，即使权利主体不同意，个人数据的传输、共享也具有合法性。

当下关口，我国正在全力实施国家大数据战略，运用大数据提升国家治理现代化水平，建立健全大数据辅助科学决策和社会治理的机制，推进政府管理和社会治理模式创新，实现政府决策科学化、社会治理精准化、公共服务高效化；推行电子政务、建设智慧城市等为抓手，以数据集中和共享为途径，推动技术融合、业务融合、数据融合，打通信息壁垒，形成覆盖全国、统筹利用、统一接入的数据共享大平台，构建全国信息资源共享体系，实现跨层级、跨地域、跨系统、跨部门、跨业务的协同管理和服务。这些战略的部署都需要政府通过广泛的样本分析了解社情民义、发展的痛点、治理的难点。毕竟，政府决策科学化、社会治理精准化、公共服务高效化都需要依赖信息的收集和利用。个人信息对于线索收集、信息溯源与情报分析的意义是巨大的，在具体制度的构建过程中需要有大局意识，处理好公权力与私权利的关系，并结合数字中国和中华民族伟大复兴的宏大背景来进行考虑。

（3）技术与法律共治的平衡

法律并不能一劳永逸地解决个人信息保护的所有问题。完全寄希望于法律制度来实现对个人信息的全天候 360 度保护，并不现实。个人信息保护需要从法律层面，结合国情明确各主体的权利与义务，并使其规则切实可行；另一方面，需要主动迎接科技进步给个人信息保护带来的全新挑战，通过技术手段设立权益保护屏障，通过创新促使保护方式变得更加智能与快捷。毕竟，新技术的应用给治理带来诸多挑战，尤其是人工智能应用使得数据采集成为常态化，区块链分布式特点会使信息暴露在大众面前，而互联网时代的个人信息保护又面临违法成本低与维权成本高的双重困境。个人信息保护需要通过技术手段来应对技术带来的难题，避免保护手段的单一化、简单化，综合运用法律、技术等多种方式求得治理效果全方位贯通无纰漏。

在实践中，有很多通过技术功能设计来保护个人信息的有益尝试。比如，电子身份证标识（eID）就是一种以密码技术为基础、以智能安全芯片为载体，由“公安部公民网络身份识别系统”签发，能够在不泄露身份信息的前提下在线远程识别身份。2017 年 10 月，全国首个将 eID 运用到不动产登记领域的项目在海口正式运行。比如，企业在研发产品时注意对个人隐私的保护，将“经设计保护隐私（PbD）”的理念贯彻产品设计之中，把隐私保护作为产品的默认设置，以预防用户隐私侵犯发生。比如，Google（谷歌）早期推出街景服务，用街景车收集街道图像，并实时展示在谷歌在线街景地图上。但谷歌街景服务却涉嫌非法收集信息、侵犯个人隐私，受到多国处罚或频频被消费者起诉。为回应规则的要求，Google 开发用户人脸和车牌虚化处理功能嵌入设计，通过技术将可识别人物身份的信息进行模糊化处理，有效地保护了个人信息。再比如，为提升用户搜索结果的准确率，Amazon（亚马逊）通常会及时存储和当前账户关联的语音搜索记录。但为了保护用户个人信息，Amazon 提供选项以供用户根据需要决定是否删除以往的搜索记录。

当下，我国正在建立网络综合治理体系，探索出一个比较好的、有利于数字中国发展的个人信息保护框架模式，应当是紧跟科技进步的步伐，导入科技驱动型的数据治理体系，坚持技术与法律共治局面。

4　结束语

数化万物，智在融通。随着产业和技术的互联互通、商业模式的创新迭代与演讲升级，数据的产生、存储、利用从消费互联网向产业互联网融合纵深延展，涉及关键基础设施的安全、产业

持续创新发展以及社会利益的重新流动。个人信息保护制度与服务百姓人格权益和财产权益息息相关，与企业创新创业和经营发展同频共振，与国家发展和社会稳定大局紧密相连。个人信息保护制度安排需要先进的立法原则和理念，需要合理的制度、科学的制度模式和具体的制度和实现路径予以保障。当然，要想制定一个符合当下文化、与经济发展水平相适应的个人信息保护制度，需要处理好个人权益保护和数据商业利用之间的平衡、公权力和私权利的平衡、法律与技术共同治理的平衡。

参考文献：

[1] 周汉华．个人信息保护法(专家建议稿)及立法研究报告[M]. 北京: 法律出版社, 2006: 79-80.
ZHOU H H. Personal information protection law (expert draft) and legislative research report[M]. Beijing: Law Press•China, 2006: 79-80.

[2] 王利明．论个人信息权的法律保护——以个人信息权与隐私权的界分为中心[J]. 现代法学, 2013, 35(4): 62-72.
WANG L M. Legal protection of personal information: centered on the line between personal information and privacy[J]. Modern Law Science, 2013, 35(4): 62-72.

[3] 张平. 大数据时代个人信息保护的立法选择[J]. 北京大学学报(哲学社会科学版), 2017, 54(3): 143-151.
ZHANG P. Legislative choice of personal information protection in the age of big data[J]. Journal of Peking University(Philosophy and Social Sciences), 2017, 54(3): 143-151.

[4] 孔令杰. 个人资料隐私的法律保护[M]. 武汉: 武汉大学出版社, 2009: 164-168.
KONG L J. Legal protection of personal data privacy[M]. Wuhan: Wuhan University Press, 2009: 164-168.

[5] 京东法律研究院. 一般数据保护条例评书及实务指引[M]. 北京: 法律出版社, 2018: 24-27.
JD Law Research Institute. General data protection ordinance commentary and practice guide[M]. Beijing: Law Press•China, 2018: 24-27.

[6] MAYER SCHONBERGER V, CUKIER K. 大数据时代[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版, 2013: 217.
MAYER SCHONBERGER V, CUKIER K. Big data: a revolution that will transform how we live, work, and think[M]. Translated by SHENG Y Y, ZHOU T. Hangzhou: Zhejiang People's Publishing House, 2013: 217.

[7] BURTON S J. 法律的道路及其影响[M]. 张芝梅, 陈刚, 译. 北京: 北京大学出版社, 2012: 268.
BURTON S J. The path of the law and its influence[M]. Translated by ZHANG Z M, CHEN G. Beijing: Peking University Press, 2012: 268.

[8] MASLOW A H. 动机与人格[M]. 许金声, 译. 北京: 中国人民大学出版社, 2007: 31.
MASLOW A H. Motivation and personality[M]. Translated by XU J S. Beijing: China Renmin University Press, 2007: 31.

[9] 龙卫球. 数据新型财产权构建及其体系研究[J]. 政法论坛, 2017, 35(4): 63-77.
LONG W Q. On the construction of new data property and its system structure[J]. Tribune of Political Science and Law, 2017, 35(4): 63-77.

[10] 张新宝．从隐私到个人信息: 利益再衡量的理论与制度安排[J]. 中国法学, 2015(3): 38-59.
ZHANG X B. From privacy to personal information: theory and institutional arrangements for re-measurement of interests[J]. China Legal Science, 2015(3): 38-59.

[11] 京东法律研究院. 一般数据保护条例评书及实务指引[M]. 北京: 法律出版社, 2018: 39.
JD Law Research Institute. General data protection ordinance commentary and practice guide[M]. Beijing: Law Press•China, 2018: 39.

[作者简介]

李美燕（1984−），女，国家计算机网络应急技术处理协调中心助理研究员，中国互联网协会个人信息保护工作委员会秘书长，北京航空航天大学法学院网络空间国际治理研究基地特聘研究员、博士，主要研究方向为互联网产业发展、个人信息保护与数据安全、互联网治理。

面向电力业务接入的跨频段融合与宽窄一体无线专网

邵炜平[1]，陆阳[2]，李建岐[2]，马平[3]，张东磊[2]
（1. 国网浙江省电力有限公司，浙江 杭州 310000;
2. 全球能源互联网研究院有限公司，北京 102209;
3. 国网绍兴供电公司，浙江 绍兴 312000）

摘　要：对电力无线专网的通信需求进行了分析，提出了面向电力业务接入的跨频段融合与宽窄一体无线专网核心理念，可以同时满足电网高速宽带、低时延高可靠、广覆盖大连接等多业务接入需求。具体提出了基于统一核心网融合方式的跨频段无线网络架构、跨频段多信道传输与基于 QCI 优先级的业务数据流调度技术以及跨频段无线网络安全防护方案。研究表明，跨频段融合无线专网兼具 LTE230、LTE1800 优势，满足电网宽带、窄带业务接入需求，相关方案的实施具有必要性和可行性，为电力无线专网的深化应用提供了支撑。

关键词：长期演进；无线专网；跨频段；融合；宽窄一体
中图分类号：TN 918.91
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018145

Cross-band fusion and wide-narrow integrated wireless private network oriented electric service access

SHAO Weiping[1], LU Yang[2], LI Jianqi[2], MA Ping[3], ZHANG Donglei[2]
1. State Grid Zhejiang Electric Power Co., Ltd., Hangzhou 310000, China
2. Global Energy Interconnection Research Institute Co., Ltd., Beijing 102209, China
3. State Grid Shaoxing Electric Power Supply Company, Shaoxing 312000, China

Abstract: The communication requirements of electric wireless private network were analyzed, and the core idea of cross-band fusion and wide-narrow integrated wireless private network oriented electric service access was proposed, meeting the needs of multiple service access simultaneously, such as high speed broadband services, low delay and high reliability services, as well as wide coverage and large connection services. Research shows that cross band fusion based wireless private network has the advantages of both LTE230 and LTE1800, and it can meet the needs of broadband and narrowband service access in the power grid. The implementation of the related schemes is necessary and feasible, which provides support for the further application of the electric wireless private network.

Key words: long term evolution, wireless private network, cross band, fusion, broadband and narrowband integrated

收稿日期：2018−01−26；修回日期：2018−04−04
基金项目：国网浙江省电力有限公司科技项目（No.5211SX16000L）
Foundation Item: Science and Technology Project of State Grid Zhejiang Electric Power Co., Ltd. (No.5211SX16000L)

1 引言

终端通信接入网是电力系统骨干通信网的延伸，负责提供配电与用电业务终端同电力骨干通信网的连接[1]。“十三五”期间，国家高度重视配用电智能化发展，要求终端通信接入网能够快速、灵活、高效地为各种业务提供支撑。智能配用电应用环境复杂、业务承载需求多样、传输可靠性要求高、终端分布区域广、测量监控点多、易受配电网扩容和城建影响。在部分场景下，光纤通信受施工难度大、建设周期长、难以全覆盖等因素制约。无线通信具有无须通信通道建设、网络部署快、系统扩展能力强等巨大的技术优势，同时，电网公司通过建设电力无线专网可以为业务传输提供可靠性、安全性保障，并可节省租用运营商无线公网的费用。电力无线专网已成为智能配用电终端通信接入网的重要组成部分。

目前，国内主流电力无线专网均采用长期演进（long term evolution，LTE）技术体制，基于230 MHz（LTE230）和 1 800 MHz（LTE1800）两种频段开展电力无线通信系统建设[2-4]。以国家电网公司为例，已先后在浙江、江苏、重庆、天津、福建等地开展了 LTE230、LTE1800 无线专网的试点应用，取得了初步成效。从应用效果来看[5,6]，LTE230 基于 223~235 MHz 电力行业授权频段，具有覆盖远、组网成本低等优势，然而由于电力行业仅获准使用该频段中非连续的 1 MHz 带宽（共 40 个信道，单信道 25 kHz），网络容量暂时不足，尚难完全满足无线专网的多业务承载需求。LTE1800 完全基于公网 LTE 技术，能够提供高带宽业务保障，具备从核心网、基站到终端的完善的产业链，然而电网公司需要单独申请 1 785~1 805 MHz 频段，单基站覆盖半径小，网络建设成本较 LTE230 偏高。总体来看，两种技术体系各有优缺点，分别适合不同的电力业务场景。

随着智能电网建设的不断推进，电力无线专网的通信需求也在快速发展。无线专网作为电网公司自有资产，如何实现一张网络同时承载电网宽带、窄带差异化业务，实现“一网多能”，成为亟待解决的问题。从目前应用情况来看，基于 LTE230 或 LTE1800 单频组网方式的无线专网尚难同时满足传输带宽、可靠性和网络覆盖范围要求，并且产品之间相互独立，接入网传输层面未能实现互联互通，使得无线网络支撑智能配用电业务的实用性受到限制。参考文献[7]对电力 LTE 异频组网系统应用进行了初步探讨，但在网络架构、业务承载与调度、网络安全防护等方面仍有待进一步明确。因此，提出充分利用 LTE230 和 LTE1800 的优势互补性，在继续发挥 230 MHz 无线频谱资源优势的同时，采用先进技术优化传输效率，深入挖掘 1 800 MHz 频段性能，建设跨频段融合与宽窄一体无线专网，实现基于统一核心网的融合以及跨频段多信道传输，满足电力高速宽带、低时延高可靠、广覆盖大连接业务安全接入需求，为电力无线专网的深化应用提供支撑。

2 电力无线专网通信需求

面向智能配用电终端通信接入网应用环境，并结合成本测算，无线专网适合于业务终端分布密度较为集中、安全性要求高的场景。按照供电区域进行划分[8]，在光纤建设困难的 A+、A 类供电区域以及 B、C 类供电区域开展无线专网建设具有较好的技术经济性。电力无线专网可以承载的业务类型主要分为电网控制类、信息采集类、移动应用类 3 种，具体描述如下。

- 部分业务通信传输容量大，实时性要求较高，如视频监控、智能营业厅业务等。单业务终端传输速率需求可达 4 Mbit/s，如果考虑业务的并发性，对带宽需求还将提高。
- 部分业务通信传输容量不大，但对实时性、可靠性要求非常高，如配电自动化、精准

负荷控制、分布式电源监控、主动配电网差动保护等。以精准负荷控制业务为例，要求端到端通信时延在毫秒级，且要求控制的安全可靠[9]；配电自动化要求实现“遥控”的可靠性，以支撑电网供电可靠性不低于 99.999%的发展目标等。

- 部分业务对实时性、传输速率要求不高，但通信数量非常庞大、信息安全需求较高，如用电信息采集、电动汽车充电桩、电网状态监测等。单业务终端传输速率需求在 kbit/s 量级，然而无线基站单扇区覆盖范围内的节点数量可达上千个。
- 部分业务对移动性、互动化通信能力要求较高，如电力资产全寿命周期管理、移动作业、移动巡检、移动营销、企业管理相关业务等。无线网络因其组网灵活性，在该类型业务承载上将扮演重要角色。

展望未来，随着智能电网和能源互联网发展，业务通信需求将进一步提升，迫切要求电力通信网能够适应以特高压电网为骨干、各级电网协调发展的新型电网模式，实现各类负荷的精准控制；引入可视化、实时化、精益化的新型作业方式，实现各级电网重要廊道的监视、巡检；开展基于“互联网+”的新型业务模式，实现用户与电网的双向互动、用电精细化管理。如图 1 所示，未来电力业务的发展，物联网业务及宽带业务并存，具有广覆盖、大连接、低时延、高可靠、高安全等特征，对电力无线专网的差异化业务支撑能力提出了更新、更高的要求。

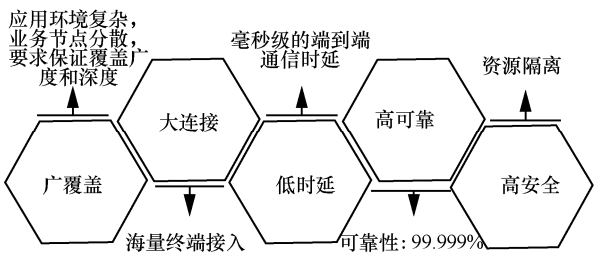


图 1　电力无线专网通信需求

3　跨频段融合与宽窄一体核心理念

LTE 是第三代合作伙伴计划（the 3rd Generation Partnership Project，3GPP）提出的“准 4G”技术，目标是实现 3GPP 无线接入技术向着更高数据速率、低时延和分组优化的无线接入技术的方向演进，包括分时长期演进（time division long term evolution，TD-LTE）与频分双工长期演进（long term evolution frequency division duplex，LTE FDD）两种方式[10-12]。目前，在 LTE 技术原理基础上发展的用于电力无线专网的无线通信系统主要采用 TD-LTE 方式，以更好地满足电力行业上下行非对称业务传输需求，具体包括 LTE230 和 LTE1800。对现有 LTE230、LTE1800 系统的相关技术参数进行比较，见表 1。

通过对比可以看出，LTE230 覆盖性能优异，但电力授权频谱资源暂时不足，对部分热点地区的宽带业务承载能力有限；LTE1800 系统满足宽带接入需求，但网络在覆盖范围、穿墙性、与业务的适配性方面存在劣势。如何在 230 MHz 基础上，结合 1 800 MHz 频段组建跨频段融合电力无线网络，在广域覆盖的同时兼顾局域热点地区的高带宽需求以及通信盲区的延伸覆盖要求，成为亟待解决的问题。结合电力无线通信需求及未来发展趋势，提出了跨频段融合与宽窄一体无线专网核心理念，具体表现如下。

所谓“宽”，有两层含义：一是系统可用的工作频段宽；二是系统能够支撑宽带业务。其中，系统可用的工作频段既包括了电力授权的230 MHz窄带频段，也包括了 1 800 MHz 宽带频段，系统能够支撑对传输速率、时延要求高的业务。目前，终端通信接入网朝着智能化的方向发展，视频监控、精准负荷控制均是其中重要的业务类型。

所谓“窄”，也有两层含义：一是兼容原有窄带 LTE230 终端及业务；二是系统能够支撑低速、广深覆盖等电力物联业务。为了充分利用已

表 1　LTE230 与 LTE1800 相关技术参数的比较

序号	技术指标	LTE230	LTE1800	总结
1	载波频段	223~235 MHz	1 785~1 805 MHz	LTE230：电力专用频点，不用再次申请 LTE1800：需要向政府部门申请频谱资源
2	工作带宽	1 MHz	5~20 MHz	LTE230：带宽小，频率离散分布 LTE1800：带宽大，频率连续分布
3	单基站峰值速率	1 MHz 带宽下，上行峰值速率 1.5 Mbit/s；下行峰值速率 0.711 Mbit/s	5 MHz 带宽下，上行峰值速率 5 Mbit/s；下行峰值速率 21 Mbit/s	LTE230：支持窄带低速率业务 LTE1800：支持宽带高速率业务
4	单基站覆盖半径	站高 45 m 情况下，城市 3~5 km，远郊 8~10 km	站高 45 m 情况下，城市 2~4 km，远郊 7~9 km	LTE230：覆盖半径相对较大 LTE1800：覆盖半径相对较小
5	时延	毫秒级	毫秒级	/
6	安全性	基于 LTE 标准加密机制，双向身份认证，采用三层安全加密体系，实现了鉴权、空口加密、非接入层信令加密和端到端加密	基于 LTE 标准加密机制，双向身份认证，采用三层安全加密体系，实现了鉴权、空口加密、非接入层信令加密和端到端加密	两者安全策略相同
7	产业链	普天	华为、中兴、鼎桥、普天等	LTE230：产业链暂时支撑较少 LTE1800：上下游产业链成熟
8	应用场景	支持窄带业务，有选择的支持多媒体传输业务	支持语音、视频和数据多媒体集群业务	LTE230：数据业务为主 LTE1800：视频、语音、数据、多媒体集群业务

建设的电力无线专网，不造成资源浪费，跨频段融合与宽窄一体无线专网必须具有前向兼容能力，兼容原有窄带 LTE230 终端及业务。此外，目前国内外掀起了低功耗广域物联网技术研究、芯片研发、模组研制的热潮，旨在实现万物互联。电力行业具有大量的数据采集、监测等具有“小数据”特征的业务，要实现真正的电网智能化，电力无线专网必须引入先进的物联网广深覆盖、低功耗相关技术，支撑电力低速业务节点的泛在互联。

在此基础上，形成跨 230 MHz 和 1 800 MHz 融合的、宽窄一体的电力无线专网，网络同时包含宽带通信终端和窄带通信终端，能够同时支持视频监控等高速宽带、精准负荷控制等低时延高可靠、采集监测等低速、广深覆盖等多样化业务终端接入，并实现和电网公司不同业务平台的对接。系统基于跨频段融合基站设备、核心网设备，采用跨频段多信道传输与业务数据流调度技术来同时承载电网差异化业务需求，从而实现电力无线专网的“一网多能”，为推动电力无线专网的深化应用指明了方向。跨频段融合与宽窄一体无线专网示意图如图 2 所示。

4　跨频段融合与宽窄一体无线专网方案

4.1　网络架构

LTE230 和 LTE1800 跨频段融合实现多频组网，可以采用紧融合和松融合两种方式。紧融合方式是指 LTE230 和 LTE1800 采用统一核心网、统一基站，通过在基站中采用不同的板卡，以使两者完全融合，仅在空口上分为 LTE230 和 LTE1800 独立天线接入，这种方式又称为统一核心网的融合方式。松融合方式是指 LTE230 和 LTE1800 采用独立核心网、独立基站，仅在上层协议实现融合，这种方式也可称为独立核心网的融合方式。

在统一核心网融合方式中，LTE230 和 LTE1800 采用统一网络层和媒体接入控制（media

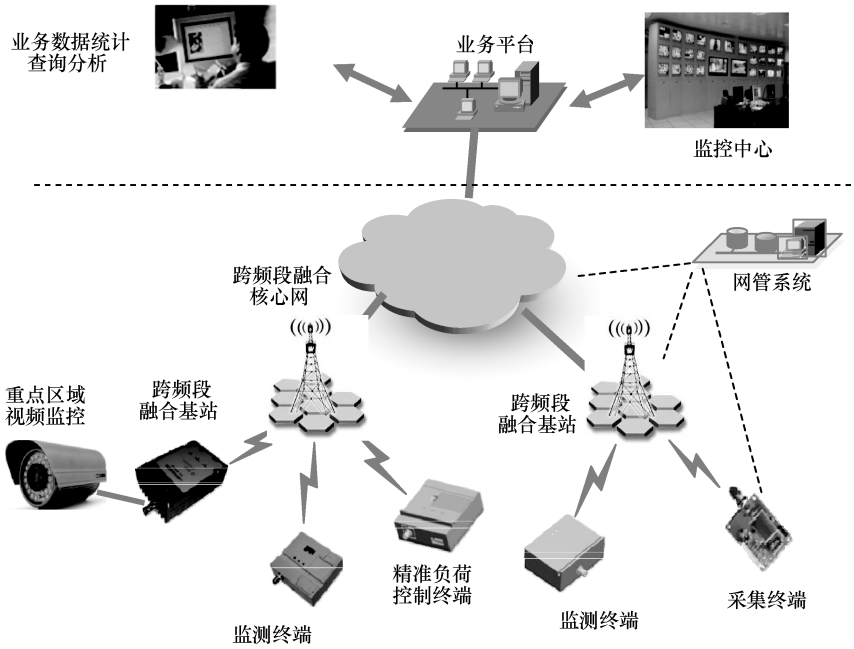
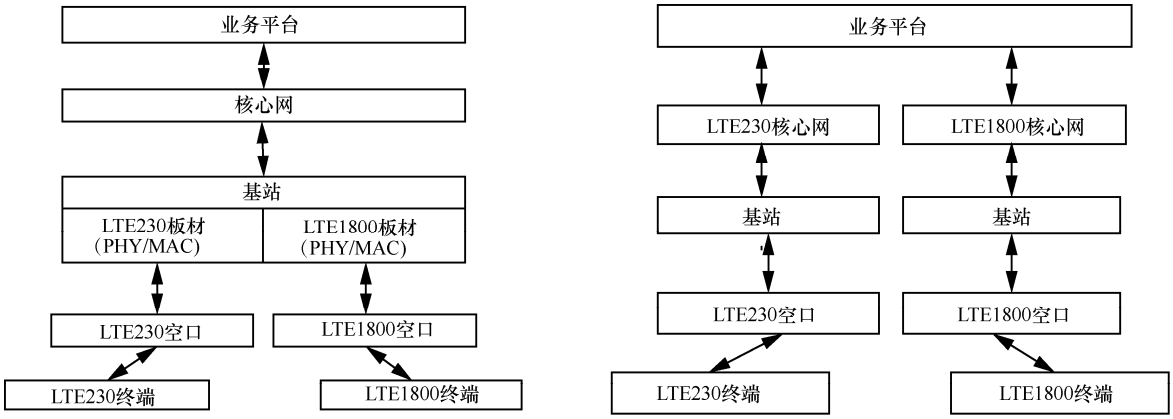


图 2　跨频段融合与宽窄一体无线专网示意图

access control，MAC）层的融合，不需要协议转换。在图 3（a）的网络架构中，将 LTE230 和 LTE1800 无线通道在 MAC 层及以上统一，但基站设计需要考虑 LTE230 和 LTE1800 无线频率、带宽等不同，采用独立的参数以及调制方式实现物理层（physical layer，PHY）。因此，基站需采用 LTE230 和 LTE1800 不同板卡的设计技术来满足二者的差异化要求。该方式下，LTE230 终端和 LTE1800 终端之间通信非常简便，不需要协议转换过程。

在独立核心网融合方式中，LTE230 和 LTE1800 采用独立核心网和基站。在图 3（b）的网络架构中，将 LTE230 和 LTE1800 无线通道在网络层及以上的协议层实现融合。该方式下，LTE230 终端和 LTE1800 终端之间通信需要进行协议的转换。



（a）统一核心网融合方式　　（b）独立核心网融合方式

图 3　跨频段融合无线网络架构

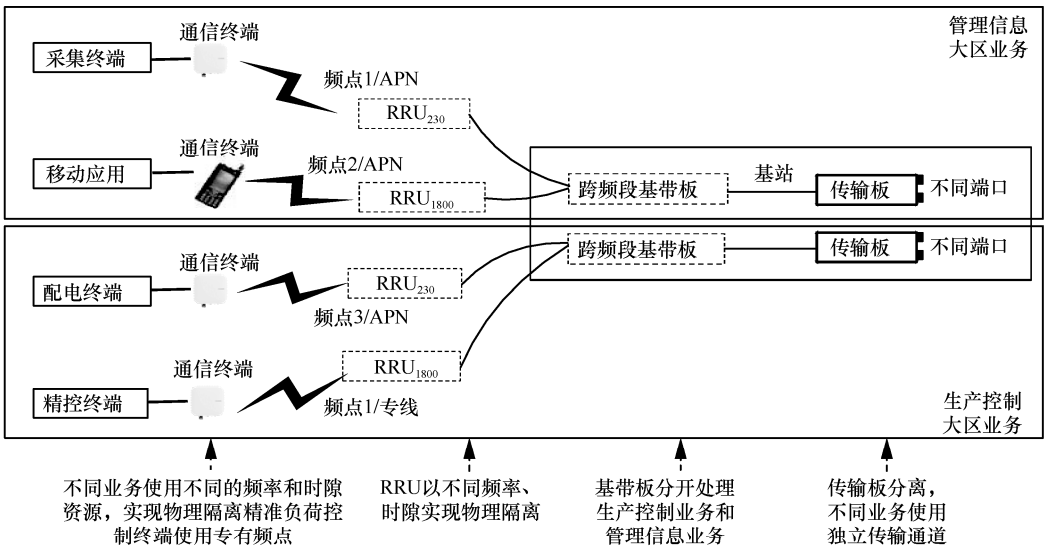
从提高业务承载效率、降低无线专网建设成本的角度出发，采用统一核心网融合方式更具备优势。因此，在设备层面，须实现跨频段融合、宽窄一体核心网，可同时接入 LTE230 基站和 LTE230 终端、LTE1800 基站和 LTE1800 终端；实现跨频段融合无线网管系统，在同一套网管上完成对 LTE230、LTE1800 多个频段无线专网的设备管理；实现 LTE230、LTE1800 跨频段融合无线基站，通过不同板卡或软件升级方式，实现对 230 MHz 和 1 800 MHz 无线射频模块的接入。跨频段融合基站基于分布式架构，包括基带单元（base band unit，BBU）和射频单元（radio remote unit，RRU），两者之间相互分离，传输的是基带信号，可以使用光纤来传输，传输距离一般在 5 km 以上。在上述统一核心网融合方式下，LTE230 以实现广域、深度覆盖为目标，LTE1800 可以在 LTE230 的覆盖范围内部署，也可以在 LTE230 的覆盖范围外部署，并重点通过 LTE1800 覆盖业务集中、数据量大的热点区域（如变电站内视频监控），从而满足差异化业务接入需求。由于部分场景下，在变电站内建设基站导致一些距离变电站较远的业务节点接入存在困难，因此可以通过在变电站部署 BBU，利用光纤将信号延伸到无线网络覆盖盲区附近再部署 RRU，从而解决网络的弱覆盖问题。对于地下室、管井等通常无线网络的弱覆盖区域，可以采用无线终端的自组织多跳级联组网方式，实现末端业务节点的灵活接入。

4.2 业务承载与调度

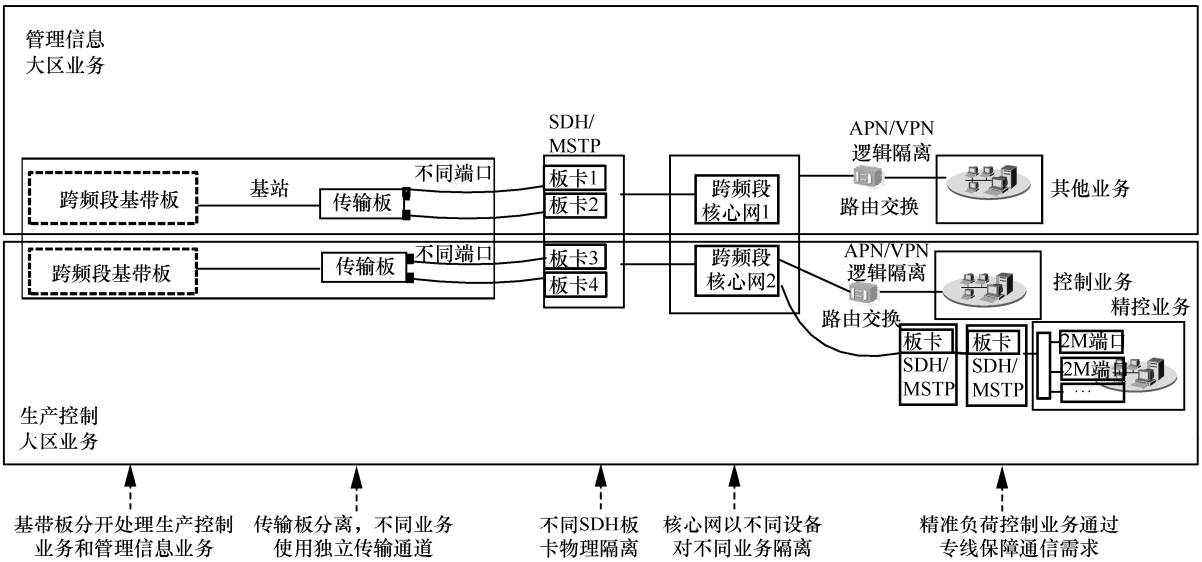
跨频段融合与宽窄一体无线专网的初衷是实现一张网络的多业务承载，因此需要按照电网公司相关规定，采取多种措施将生产控制大区与管理信息大区业务传输通道进行物理隔离[13]。提出跨频段融合与宽窄一体无线专网通过空口不同的频率、时隙资源，基站双跨频段基带板、传输板/端口，双跨频段核心网/板卡和物理隔离的同步数字序列（synchronous digital hierarchy，SDH）/多业务传送平台（multi-service transport platform，MSTP）通道为不同的业务提供专线通道，实现端到端的跨频段多信道传输，并通过接入点名称（access point name，APN）/虚拟专用网（virtual private network，VPN）实现管理信息大区内业务之间的逻辑隔离。

如图 4 所示，作为一个例子，采集业务终端、移动业务终端、配电业务终端、精控业务终端分别通过不同的无线终端以及 RRU 接入，其中，RRU_{230} 代表 LTE230 的射频单元，RRU_{1800} 代表 LTE1800 的射频单元，并通过可以灵活配比的、不同的空口接入频率或时隙，实现生产控制大区与管理信息大区业务的物理隔离。基站采用双跨频段基带板、传输板/端口，并尽可能利用变电站等电网公司自有物业进行建设。核心网实现在市公司通信机房的地市部署，不同的核心网设备/板卡分别承载生产控制大区和管理信息大区业务。其中，针对服务质量（quality of service，QoS）要求高的精准负荷控制业务，为其分配专用的时域、频域资源，实现专线通道传输，其他业务则采用共享网络接入方式，由系统提供 QoS 优先级保障。提出 LTE230、LTE1800 射频单元均可接入电网控制类终端（如配电自动化“三遥”、精准负荷控制），实现更快速、更安全的控制；用电信息采集、电网状态监测等信息采集类终端重点由 LTE230 来接入，满足未来信息采集量更多、频次更高的要求；移动作业、移动营销等移动应用类终端重点由 LTE1800 来接入，满足电网公司各专业、随时随地支撑现场业务（特别是视频监控类高带宽业务）的要求。

跨频段多信道传输离不开高效业务调度技术的支撑。提出基于业务 QoS 参数（QoS class identifier，QCI）优先级实现业务数据流调度，通过为不同业务配置 QCI 优先级，针对 QCI 优先级提供网络传输资源保障。首先，根据不同电力数据业务对时延、传输速率、可靠性的不同需求，



(a) 无线网（空口至基站）跨频段多信道传输



(b) 回传网（基站至核心网）跨频段多信道传输

图 4　跨频段多信道传输实现多业务承载

对网络所承载的电力数据业务进行优先级划分。基站调度时，根据业务 QCI 优先级来分配时隙、频率资源，优先保证控制类低时延高可靠业务接入。当检测到拥塞时，丢弃位于严重拥塞阈值后的低优先级业务，一方面优先保证了高优先级业务的 QoS 需求，另一方面缓解了拥塞节点的拥塞程度。在为被丢弃的低优先级业务重新建立传输通道时，综合考虑备选路径的时延与可用带宽，不仅可以保证业务的 QoS，且优化了网络资源，均衡了整个无线网络的负载。

4.3　网络安全防护

面向智能配用电终端通信接入网的跨频段无线专网可同时承载生产控制大区、管理信息大区业务，导致网络跨接在物理隔离的两张信息内网之间，可能引起两张网络的物理隔离被打破，造成潜在的安全风险，因此，网络安全防护是不可

或缺的一环。其中，配电环节的业务处于生产控制大区，直接服务于生产调度，对安全防护等级要求较高。按照相关标准，为了保证控制的安全可靠性，具备遥控功能的业务应优先采用专网通信方式，保证一次设备的安全运行，如采用无线专网实现信息接入应符合相关安全防护规定，并有严格的安全防护策略。与此同时，随着分时电价、阶梯电价、远程费控等用电营销业务需求出现，用电环节对通信安全性的要求进一步提高，电价和用电信息的传送必须要有安全、稳定、可靠的通信通道。在采用电力通信专用通道传输此类信息时，网络信息安全要在原有用电信息采集系统的基础上达到更高的层次。综上所述，根据电力通信网安全防护总体要求，电力通信接入网安全接入方案应根据不同安全分区特点，对生产控制大区和管理信息大区分别设计防护措施，并重点在主站侧、终端侧和边界处予以部署[13]，其总体框架如图 5 所示。

结合上述分析，并借鉴 4G 无线公网的安全防护机制以及 LTE 无线专网可采用的安全防护策略，提出面向跨频段融合与宽窄一体无线专网的安全防护方案，主要包括：

- 空口以不同的时隙、频率实现不同业务的物理隔离；
- 跨频段融合基站以不同的基带板卡实现不同业务的物理隔离；
- 跨频段融合基站与核心网间采用 SDH/MSTP 单独通道或网际安全协议（internet protocol security，IPSec）加密传输，优先考虑使用 SDH/MSTP 单独通道，不同业务的 SDH 板卡实现物理隔离；
- 跨频段融合核心网以不同设备/板卡对不同业务进行隔离；
- 通信终端与安全接入分区接入设备之间采用 VPN 进行逻辑业务通道隔离，加强安全加密措施；
- 业务终端 MAC 地址绑定，防止合法终端被窃取后，被不法人员用于攻击网络；
- 通信终端可采用国际移动用户识别码（international mobile subscriber identity，IMSI）绑定措施，采用 eSIM 形态，消除终端被盗用、复制的可能性；
- 对终端与基站之间数据进行加密传输、完

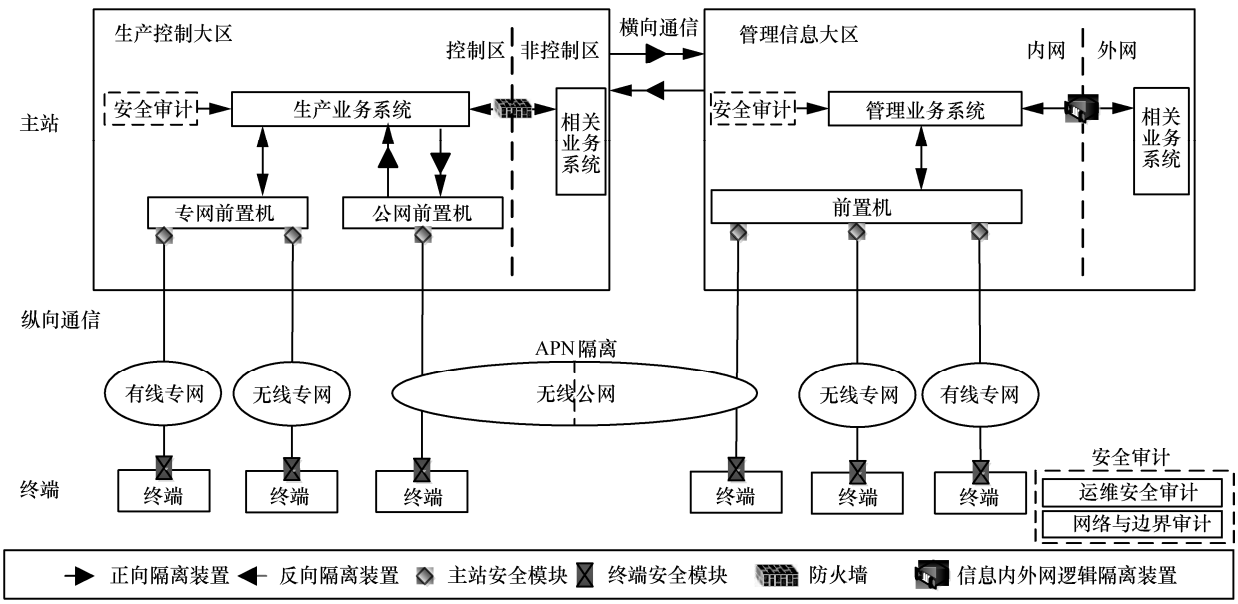


图 5　电力通信接入网安全接入方案总体框架

整性保护，防止空口数据被截获，实现信息隐蔽；

- 对终端与核心网之间采用非接入层（non-access stratum，NAS）加密、双向认证鉴权、完整性保护，防止伪基站窃取、篡改终端数据，伪终端入侵网络以及高层控制信令被截获和篡改；
- 通过以上措施，对跨频段融合与宽窄一体无线专网承载电力业务接入，特别是配电自动化“三遥”、精准负荷控制等电网控制类业务提供有效的安全防护。

5　结束语

LTE 无线专网是智能配用电终端通信接入网的重要组成部分，然而现阶段基于230 MHz、1 800 MHz单频组网方式建设的 LTE 无线专网尚难同时满足传输带宽、可靠性和网络覆盖范围要求。结合电力无线专网通信需求分析，指出未来电网业务将呈现高速宽带、低时延高可靠、广覆盖大连接等差异化属性。以业务发展趋势为牵引，提出了面向电力业务接入的跨频段融合与宽窄一体无线专网核心理念，形成了跨 230 MHz 和 1 800 MHz、从核心网层面深度融合的无线网络架构。提出了230 MHz 和 1 800 MHz 具备业务覆盖侧重、生产控制大区业务与管理信息大区业务传输通道端到端物理隔离的跨频段多信道传输方案，基于 QCI 优先级的业务数据流调度技术以及符合电力通信接入网安全接入总体要求的跨频段无线网络安全防护方案，为电力无线专网的深化应用与不断演进提供了支撑。

参考文献：

[1] 国家电网公司信息通信部. 终端通信接入网统筹建设总体设计[R]. 2017.
Department of Information and Communication, State Grid Corporation of China. Overall design of the construction of terminal communication access network[R]. 2017.

[2] 欧清海, 谢杰洪, 曾令康, 等. TD-LTE 技术在配用电通信中的应用[J]. 现代电子技术, 2012, 35(23): 27-31.
OU Q H, XIE J H, ZENG L K, et al. Application of TD-LTE technology in power distribution and utilization system[J]. Modern Electronics Technique, 2012, 35(23): 27-31.

[3] 周建勇, 田志峰, 李艳, 等. 广覆盖 LTE230 系统在电力配用电应用中的研究与实践[J]. 电信科学, 2014, 30(3): 168-172.
ZHOU J Y, TIAN Z F, LI Y, et al. Research and practice of LTE 230 system with wide coverage characteristics in the power distribution and utilization application[J]. Telecommunications Science, 2014, 30(3): 168-172.

[4] 徐杰, 侯功华, 何尚骏, 等. 基于 TD-LTE 1 800 MHz 的电力无线专网覆盖性能优化研究[J]. 信息通信, 2018, 16(1): 127-128.
XU J, HOU G H, HE S J, et al. Research on coverage performance optimization of power wireless private network based on TD-LTE 1800 MHz[J]. Information & Communications, 2018, 16(1): 127-128.

[5] 蔡根, 张健明, 杨大成. TD-LTE 电力专网 230 MHz 与 1.8 GHz 的研究[J]. 软件, 2015, 36(12): 83-88.
CAI G, ZHANG J M, YANG D C. Research on TD-LTE 230 MHz and 1.8 GHz telecommunications network for electric power[J]. Computer Engineering & Software, 2015, 36(12): 83-88.

[6] 闫淑辉, 冯世英. 现有无线宽带通信技术在电力行业应用对比分析[J]. 计算机科学与探索, 2016, 10(S1): 670-675.
YAN S H, FENG S Y. Comparison and analysis of existing wireless broadband communication technology in power industry[J]. Journal of Frontiers of Computer Science and Technology, 2016, 10(S1): 670-675.

[7] 姚继明, 黄凤, 田文锋. 电力 LTE 异频组网系统应用研究[J]. 计算机技术与发展, 2017, 27(3): 181-184.
YAO J M, HUANG F, TIAN W F. Research on application of anti-frequency networking of LTE for power system[J]. Computer Technology and Development, 2017, 27(3): 181-184.

[8] 国家电网公司. 配电网技术导则: Q/GDW 10666[S]. 2016.
State Grid Corporation of China. Technical guidelines for distribution network: Q/GDW 10666[S]. 2016.

[9] 国家电网公司信息通信部. 关于印发精准负荷控制通信系统建设指导意见（试行）的通知[S]. 2017.
Department of Information and Communication, State Grid Corporation of China. Notifications for the guidance of the construction of the precision load control communication system (trial)[S]. 2017.

[10] 李文宇, 宋丽娜, 何秀淼. LTE 产业发展分析和展望[J]. 电信科学, 2014, 30(3): 6-11.
LI W Y, SONG L N, HE X M. Development analysis and perspective on LTE current status[J]. Telecommunications Science, 2014, 30(3): 6-11.

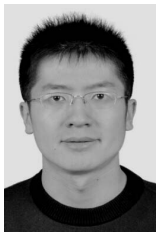
[11] 王伦锁. TD-LTE 与 LTE FDD 融合组网策略[J]. 电信科学, 2016, 32(1): 188-192.
WANG L S. TD-LTE and LTE FDD fusion networking strategy[J]. Telecommunications Science, 2016, 32(1): 188-192.

[12] 庞晓丹, 李薇薇, 孙茜, 等. LTE 无线网络虚拟化中切片调度策略[J]. 电信科学, 2017, 33(2): 66-72.
PANG X D, LI W W, SUN Q, et al. Slice scheduling strategy in LTE wireless network virtualization[J]. Telecommunications

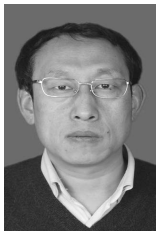
Science, 2017, 33(2): 66-72.
[13] 中华人民共和国国家发展和改革委员会. 电力监控系统安全防护规定[S]. 2014.
People's Republic of China National Development and Reform Commission. Safety regulation of power monitoring and control system[S]. 2014.

[作者简介]

邵炜平（1976−），男，国网浙江省电力有限公司高级工程师、通信处副处长，主要从事基于电网业务需求的通信应用技术研究工作。

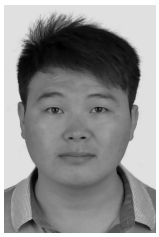


陆阳（1984−），男，博士，全球能源互联网研究院有限公司高级工程师，主要从事无线通信、电力线载波通信及其在智能电网中的应用研究工作。



李建岐（1969−），男，全球能源互联网研究院有限公司高级工程师（教授级）、信息通信研究所副总工程师，主要从事电力通信技术研究及开发工作。

马平（1962−），男，国网绍兴供电公司高级工程师，主要从事电力系统通信网络规划建设和运行管理工作。



张东磊（1986−），男，全球能源互联网研究院有限公司工程师，主要从事电力通信技术应用研究工作。

互联网跨域端到端质量监测及故障定位方案

颜永明[1]，陈兵[2]，许文杰[1]
（1. 中国电信股份有限公司上海分公司，上海 200085;
2. 上海市信息网络有限公司，上海 200081）

摘　要：随着互联网市场的迅猛发展，互联网内容提供商对网络质量提出了更高的要求。大型互联网内容提供商因业务需要，应用遍布于各地数据中心，对互联网跨域访问质量有很高的要求。建立互联网跨域端到端质量监测系统，快速定位域外网络故障，对互联网运营商及服务提供商具有较大的意义。分析了常用网络监控技术和等价路径中的散列（Hash）算法，提出利用散列算法实现网络路径全遍历，监测互联网跨域端到端质量，最后对故障定位方案给出了建议。

关键词：跨域；端到端质量；散列；路径全遍历
中图分类号：TN915.41
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018239

Internet cross-domain end-to-end quality monitoring and trouble location scheme

YAN Yongming[1], CHEN Bing[2], XU Wenjie[1]
1. Shanghai Branch of China Telecom Co., Ltd., Shanghai 200085, China
2. Shanghai Information Network Co., Ltd., Shanghai 200081, China

Abstract: With the rapid development of internet, the internet content providers need much better quality of internet. They usually need to deploy their applications in the internet data centers all over the country due to their business requirements. Users usually cross domain to access in their data centers. So internet content providers have high level of requirements for cross-domain internet quality. It is very important for internet service providers to establish a monitoring system which could monitor end-to-end cross-domain internet quality and locate the troubles of network out of autonomous systems rapidly. The commonly used network monitoring technology and the hash algorithm in equal-cost multipath routing were analyzed. A solution of using Hash algorithm to realize full traversal of network paths and monitor end-to-end cross-domain internet quality, was presented. Some suggestions were also given for trouble location scheme.

Key words: cross domain, end-to-end quality, Hash, full traversal of network paths

1　引言

在各类移动端和桌面端应用，如游戏、社交应用、购物以及搜索引擎等需求的推动下，互联网企业对网络带宽的需求持续增长，数据中心经多年发展，建设规模不断扩大。同时 ICP（internet

收稿日期：2018−06−21；修回日期：2018−08−06

content provider，互联网内容提供商）对网络访问质量提出了更高的要求。云计算数据中心更因业务跨机房调度、配置部署等特性，对网络高可靠性、稳定性、网络时延、分组丢失、抖动等质量问题更为敏感[1]。网络异常对这些业务的用户体验造成极大影响。因此，本文针对因自治域问题而无法直接采集到网元、链路状态的互联网跨域场景，重点研究了端到端质量监测及故障定位方案。

2 互联网跨域端到端质量监测场景需求和问题

目前，大型互联网公司大多在全国租用或自建数据中心，网络游戏、社交应用、在线购物及搜索等各类业务分布式部署，访问这些服务的用户来自全国各地，且数量庞大。以手机端游戏业务为例，玩家来自全国各地，而服务器则可能集中部署于北京、上海或者广州等主要城市的数据中心，玩家使用手机终端经本地运营商接入网络、城域网、骨干网，再到业务所在数据中心城域网，数据中心服务器经返程完成完整的数据交互。此类业务特性决定了 ICP 非常关注跨域端到端网络质量。从运营商收到的日常申告来看，ICP 主要反映从数据中心到某个/某些省市的访问质量劣化，影响业务，并经常要求确认相应省间、区域间（例如全国往华东区域）的网络是否存在问题。事实上，只要端到端路径中某个段落的网络出现异常，就有可能会导致分组丢失、拥塞、时延增大、数据重传，造成游戏卡顿，降低用户体验[2]，因此，亟需建立一套端到端网络质量监控系统来实时监控网络质量，快速定位故障点，以便 ISP（internet service provider，互联网服务提供商）/ICP 及时排障、迂回，确保业务质量保持稳定。

3 传统端到端质量监测和故障定位方式

传统网络质量监控通过部署在数据中心的网络探针向目的被监控 IP 地址进行单向 ping 和 trace 测试，以发现、定位故障[3]。这一方式存在如下缺陷：

- 探针部署位置固定，无法拟合用户数据分组转发路径；
- 由于是网络探针单向做 ping 和 trace，当监测到故障时无法判断是 ICMP（internet control message protocol，Internet 控制报文协议）数据分组去程方向路径异常还是回程路径异常；
- 目前数据网络都具有冗余保护措施，网络设备都是多上联架构，每条路径都有 2 个以上的冗余节点，整个网络路径呈现复杂的 ECMP（equal-cost multipath，等价多路径）路由，传统的单向监测的探针 IP 地址和目的 IP 地址相对固定，导致监测路径也是固定的，由于无法遍历每条路径，若监测数据分组正好走在状态正常的路径，就无法监测其他可能发生质量劣化的等价链路，大概率与实际用户业务数据分组流经的路径不一致，对网络质量监测和故障定位造成严重干扰[4-5]。

4 互联网跨域场景下网络路径全遍历的质量监测方案

4.1 端到端网测监控系统技术分析

有多种方法可实现支持目前主流网络协议场景的网络质量监控，分为被动式和主动式网络质量监测。

被动式网络质量监控，如网元告警系统等是运营商常用的监控方式。该监控室采集各种信息例如设备告警、端口流量等，然后把采集到的信息传送到相关服务器进行筛选、分析、告警和存储等处理。但考虑到用此种方式采集到的数据主要是监控自治域内的网络设备、链路告警，存在很大的局限性，且由于维护权域问题，很难做到跨域的端到端网络质量监控。

采用主动式网络质量监控，在网络部署探针发起监测，对网络上的一些目的地址，例如实际用户访问的互联网网站，也可以通过数据中心间的网络探针做双向监控，根据应用场景灵活部署监控任务。

IP 在当前网络中使用最广泛，ICMP 作为 IP 的一个子协议具有一定的控制能力和网络监控能力，经常用于网络监控和链路故障定位。

4.1.1　ping 测试

ping 测试通过发送 ICMP 数据分组给目标主机，要求目标主机返回应答消息。ICMP 是 TCP/IP（transmission control protocol/internet protocol，传输控制协议/互联网互联协议）协议族的一个子协议，属于网络层协议，广泛使用于检测网络连通性、路由可用性、目的主机是否可达等。遇到数据分组无法到达目的网络、网络带宽速率低、时延高等情况发生时，可发送 ICMP 消息，测试当前网络状态。

ping 测试将 ICMP 数据分组发出，目的 IP 地址对收到的数据进行分组后检查目的地址，如和自身 IP 地址相符则接受，并且把数据分组中相关信息交给 ICMP 处理，封装应答数据分组后回传给源地址。当 ping 测路径异常时，返程应答数据分组里包含路由不可用、目的主机不可达等错误信息。在日常网络运维中经常使用 ping 测试用于检测网络连通性。

4.1.2　trace 工具

trace 发出 TTL（time to live，存活时间）值追踪到达目的主机所经过的节点，从节点收到 ICMP 检测的应答数据，用来检测数据发送源到目的节点之间所经节点。信息在网络中传输会经过服务器、交换机、路由器等网络设备，这些设备通常都配置 IP 地址，trace 可以让运维人员知道数据端到端传输所经过的路径。trace 可以测量数据分组从发送到目的地址再返回源地址所需要的时间，其测试报告包含节点 IP 地址、每个节点所花费时间、分组丢失率等信息。

需要注意的是，根据散列算法数据分组从源地址到目的地址走的路径可能会随着每次测试而改变，使用固定源、目的 IP 地址测试分组测试网络状态时，ping、trace 报告显示路径正常，只能代表这次 ping、trace 测试分组恰巧经过了正常路径，但如果用户业务数据分组经过的是等价链路中的劣化路径，则测试结果与用户感知不一致。此外，每一 trace 返回数据分组因源地址被修改为响应设备地址，五元组散列结果差异会造成回程路径的不同，因此测试结果无法真实反映实际情况。

4.2　散列算法实现全路径覆盖监控方案

数据分组在等价路由转发时，网络设备会将每个连接的所有数据分组都发往多个链路中的其中一条，数据分组建立连接和转发链路是通过网络设备散列算法来完成的，散列算法是根据散列因子来计算路径的[6]，比如，包含源 IP 地址、源端口、目的 IP 地址、目的端口和传输层协议的五元组。

发起监控 ICMP 分组时的散列因子和用户业务数据流建立连接时的散列因子不同，用户数据流恰好走故障链路而影响业务，而监控流经过的链路则有可能是正常路径。日常运维中用户业务有时延、分组丢失等现象，而 ICMP 监控报告则显示正常，此时就有可能是在等价链路中监控数据分组是通过链路 A 传输，而用户业务数据分组是从链路 B 转发。这就要求网络质量监控系统首先要满足监控路径覆盖全链路，实现遍历性。

4.2.1　单向监测

部署在数据中心的探针发起对外部公网互联网网站的 ICMP 监控。探针支持多 IP 地址配置，选取多个互联网网站作为目的 IP 地址进行监测。探针发起对不同目的互联网网站的 trace，探针切换不同源地址继续对不同互联网 IP 地址发起测

试，部署在不同网段的其他探针也持续重复以上监控。经过上述监控步骤可以得到数据中心到互联网的不同访问路径，画出全路径网络监测拓扑结构，如图 1 所示，可以看出左边探针发起对互联网网站地址的 ICMP 监控，当探针用不同源地址或者多个探针发起监控时，会经过多个不同中间节点，满足监控链路的遍历性要求。逐渐提高目的互联网网站地址的数量，查看中间路径层数和节点数是否会明显增加，如果调整到无明显增加时，可以得到覆盖整个中间路径的网络监控拓扑。

单向监测的特点是每个探针进行一对多监测，目的 IP 地址可以在探针最大性能范围内找海量公网存活 IP 地址，根据散列算法，足够多的源–目的 IP 地址因子组合可以实现监测覆盖全路径的效果，用少量探针实现监测遍历性，可以减少部署在异地数据中心的探针数量，控制成本。而缺点是单向监测只能从全局宏观上发现某一方向的网络质量异常情况，定性给出网络变化情况，无法精确定位故障点。

4.2.2 双向监测

用户网络服务是用户发送请求到数据中心，再从数据中心反馈内容，是双向的，因此，网络设备在两个方向分开进行散列算法。如果说用户到数据中心的方向是路径 A，那么数据中心分组返回时走的路径可能是 B，而 ping 和 trace 报告都是单向监控，如果需要监控双向路径，就要在正反方向分别相互做 ping 和 trace 监控。单向监控的局限性决定，单向监控只能用来做告警发现和绘制全路径网络监控拓扑图，要实现故障点定位功能，需要结合探针双向监控。

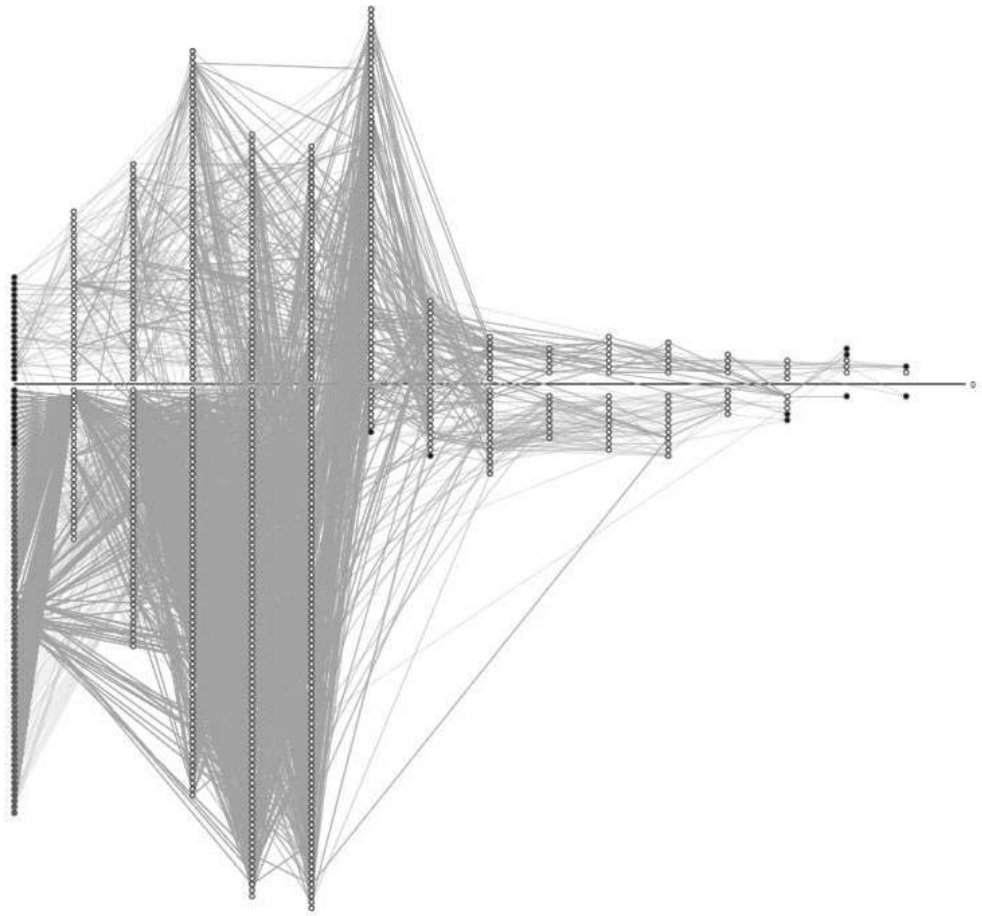


图 1 数据中心到公网网络监控拓扑结构

在日常网络运维中会碰到以下情况，见表 1。

如表 1 所示，正向 trace 报告中第 10 跳已到用户侧网络且显示分组丢失出现在最后一跳。第 10 跳分组丢失率为 0，依据排障经验可以判断之前显示的分组丢失率为中间节点设备对 ICMP 分组做的保护机制，对正常业务数据转发无影响，故障点在最后一跳用户设备。

但是从表 2 反向 trace 报告可看出，分组丢失出现在第 8 跳节点，之后分组丢失率一直延续，故障点在第 8 跳省际链路中间节点。最后的排查结果是上海到北京的一根链路故障导致的拥塞引起分组丢失。业务请求和数据反馈走的双向路径会不同，且故障往往发生在其中一条链路上，在实际日常运维中会经常碰到结合单向和反向监测报告判断出故障点的情况。

在需要监控路径的多个数据中心部署多个探针，探针支持多 IP 地址配置，数据中心中的多个探针与另外数据中心探针进行相互 ICMP 监控。不同数据中心探针间形成 full-mesh 监控，如图 2 所示。

探针源地址数和探针部署量，可以根据单向监控方式绘制的全路径网络监控拓扑图作调整。探针间 full-mesh 监控结果可得出数据中心间全路径覆盖监控，且可以准确定位路径中的故障节点，如图 3 所示。

4.3　端到端网络质量监控方案的实现

系统由网络探针和数据处理服务器组成，网络探针用于发起监控，收集监控数据，数据处理

表 1　正向 trace

IP地址	数据分组				ping		
	分组丢失率	次数	最近	平均	最佳	最差	标准偏差
1．123.1xx.xx.x	0	138	0.6	0.8	0.5	3.3	0.3
2．???							
3．123.15x.xx.xx	52.6%	138	2.6	2.5	1.6	5.8	0.7
4．221.2xx.xx.xxx	0	138	1.8	2.0	1.6	10.0	0.9
5．221.2xx.x.xxx	0	138	1.7	1.9	1.6	10.7	1.1
6．201.9x.xx.xx	59.1%	138	4.7	4.5	4.4	4.7	0.1
7．101.9x.xx.xxx	95.6%	137	32.1	33.3	32.0	39.8	3.1
8．???							
9．101.xx.xxx.xxx	29.4%	137	28.5	28.5	28.4	28.7	0.1
10．101.2.xx.xxx.xx	0	137	29.7	29.6	29.3	30.0	0.1
11．???							
12．101.xx.x.xx	0.7%	137	131.1	130.6	127.6	131.4	0.7

表 2　反向 trace

IP地址	数据分组				ping		
	分组丢失率	次数	最近	平均	最佳	最差	标准偏差
1．101.xx.x.x	0	119	0.7	0.7	0.5	1.5	0.2
2．???							
3．101.2xx.xxx.xx	41.2%	119	0.7	0.7	0.6	1.9	0.2
4．101.9x.xxx.xxx	0	119	1.4	1.7	1.4	8.7	0.9
5．101.xx.xxx.xxx	0	119	6.3	6.0	1.8	9.6	2.4
6．202.xxx.xx.xxx	0	119	128.4	129.9	125.6	139.5	1.6
7．202.xx.xx.xx	0	119	130.5	131.3	127.3	133.9	1.5
8．221.2xx.x.xxx	44.1%	119	130.7	130.3	125.9	131.1	0.9
9．???							
10．12x.xxx.xx.xx	0.8%	119	130.8	130.4	126.9	131.3	0.8
11．???							
12．12x.xxx.xx.xx	2.5%	118	130.9	130.5	127.1	131.2	0.a7

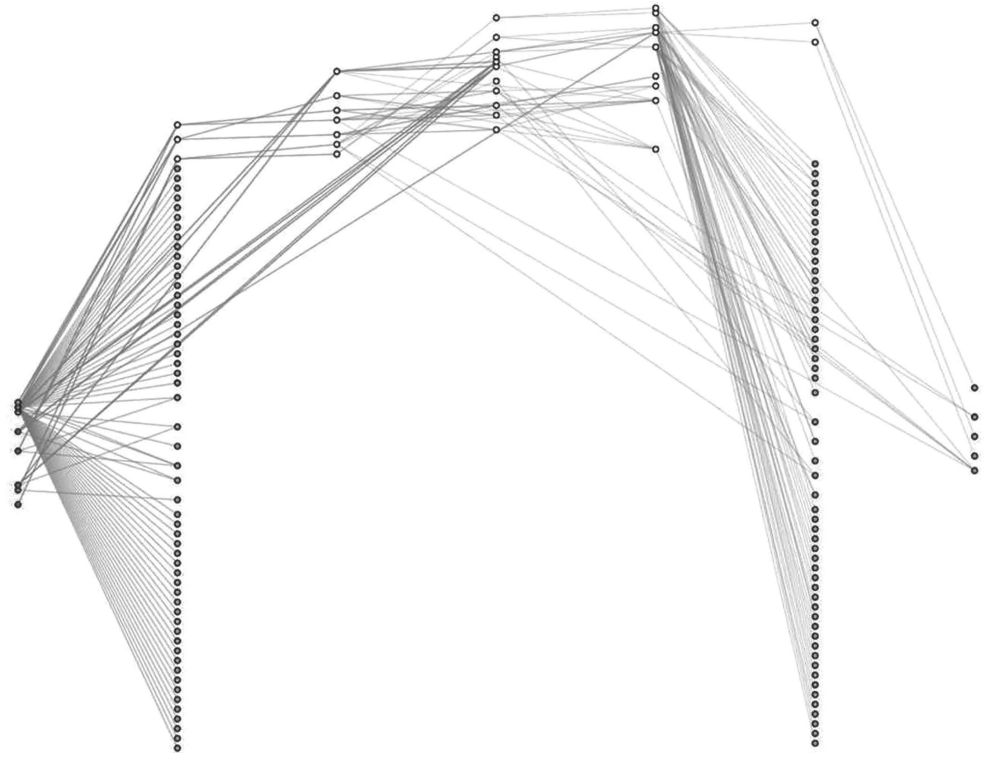


图 2　数据中心间形成 full-mesh 监控

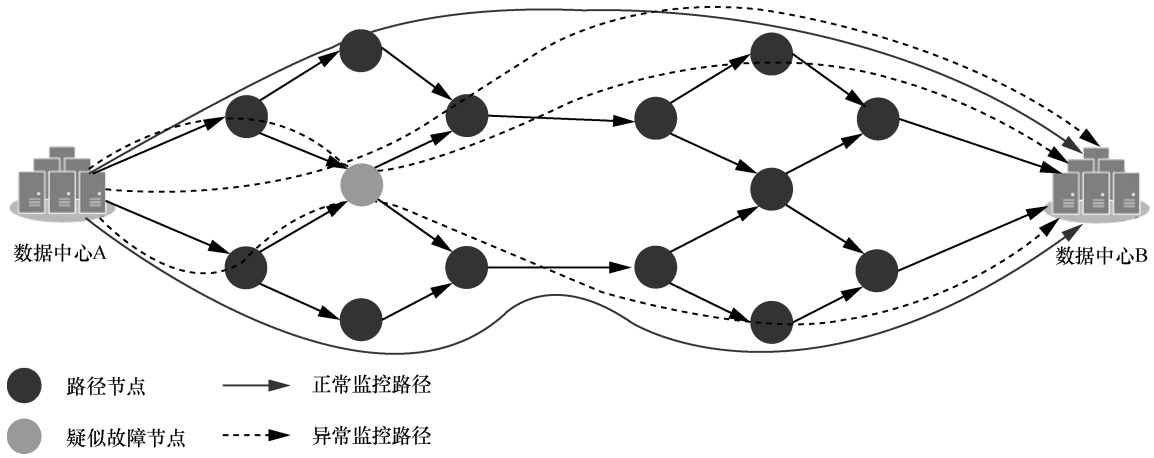


图 3　定位故障点

服务器负责对探针进行配置下发和管理，然后对收集到的监控数据进行处理。方案中单向监测和双向监测互为补充。

单向监测选取海量互联网网站地址作为目的 IP 地址，可以用少量探针达到监测遍历性的效果，用单向 trace 可以画出某方向的全路径监测拓扑，并且根据周期性监测数据动态调整拓扑图，同时，用单向 ping 方式得到某一方向的宏观网络质量。双向监测在满足遍历性的情况下实现故障点精确定位，通过两端数据中心部署的探针相互 ping/trace，可以做到快速发现故障并准确定位故障点。

从单向和双向监测的历史数据中可以看出网络周期性变化情况，分析计算出各方向、各时间段的网络质量基线，再将网络质量基线应用在日常实时监测中，当实时监测数据偏离质量基线时

触发告警。为了减少探针和网络负荷，在日常监测中单向和双向监测频率可以适当降低，且只进行 ping 测定性网络质量情况。当定性发现异动时再触发高频率的监测，用双向 trace 方式定位故障点。

探针采用虚拟化方式部署在云资源池。虚拟化部署成本低、部署简便，其次在后续维护中无需到达现场，在远端即可登录查看，方便管理。因为探针部署在省外，要确保探针不会失联。虚拟化探针支持配置多个 IP 地址，当数据上传或者管理通道走在故障路径时，只要探针本身状态正常没有僵死，或者链路全断的情况外，数据处理服务器可以通过轮询配置好的多个 IP 地址正常工作。

通过端到端的应用监控，可以很快掌握网络中间的故障点和用户体验，进一步迅速排查网络出现的问题、定位故障点，为减少故障和提高用户体验，起到了非常重要的作用。

4.3.1　了解网络拓扑变化

通过网络探针持续监控各个方向的网络质量和监控路径，可以掌握整个网络拓扑信息。根据历史监控数据，周期性调整拓扑图，比如一周更新一次拓扑图。根据拓扑信息，了解网络变化的趋势，还可以为以后的网络变更及带宽扩容等长期规划提供参考依据，对某些经常发生网络故障的链路和节点及时做优化和整改方案。

4.3.2　制定网络性能基线

长期监控链路质量，根据节点性能，将忙时和闲时、工作日和节假日等不同维度的网络质量监控数据形成动态调整的基线。监控系统把实时收集的监控数据和告警基线相比较分析，之后弹出故障点、故障等级等告警信息，提醒运维人员确认，后续跟进处理。

4.3.3　收敛链路告警数据

数据处理服务器根据链路方向、网络拓扑数据等信息来确定网络探针收集到的监控数据。例如，同一路径方向上的告警数据归为一条告警，网络拓扑数据中属于相同网络层级或同一网络设备的 IP 地址告警归为相同告警，避免运维人员在海量告警数据中浪费故障处理时间。根据不同监控场景，制定多种收敛算法。

4.3.4　多维度报表

可以根据网络质量监控数据多维度提供网络性能报表，根据不同区域、不同数据中心、不同访问方向、多个监控路径等维度来掌握网络质量信息。还可以根据新的监控场景和业务场景，自定义质量监控数据，以图表和报表方式呈现网络健康状态。

4.4　网络监控拓扑

根据需求该系统可以实现的监控网络拓扑如图 4 所示。

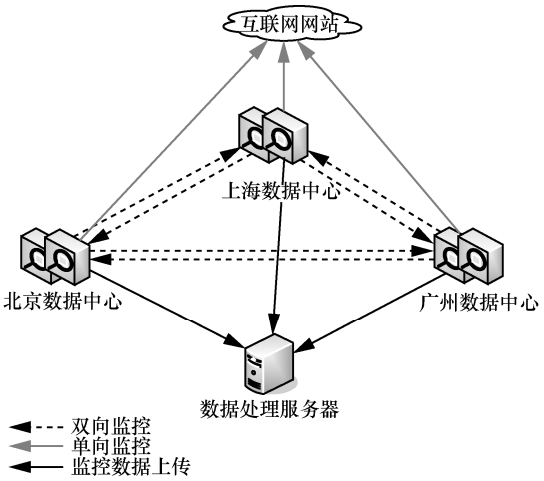


图 4　监控网络拓扑

具体如下：

- 在主要城市数据中心部署网络探针；
- 中间数据处理服务器为部署在各省市数据中心的网络探针下发配置、任务，集中管理管理部署；
- 部署在各省市数据中心的网络探针相互形成 full-mesh 监控数据流，进行双向监控；
- 部署在各省市数据中心的网络探针也发起对互联网网站 IP 地址的监控数据流，进行单向监控。

4.5 与传统网络质量监测方式对比

数据分组在网络中传递时会经过多个节点，现在网络为了冗余性，节点和节点间用多个等价链路做负载均衡，通常会用散列算法来实现。使用五元组（源地址、目的地址、源端口、目的端口、协议类型）为散列因子，经过散列算法得出的散列值作为区分不同数据流的标志，映射到不同等价链路。在 ICMP 测试中，每次 ping 测试或者 trace 都计算为一次 ICMP 连接，所以每次 ICMP 测试都只会监控多条链路中的一条，在此次 ICMP 测试中其他等价链路不会被监控到。发起监控 ICMP 分组时的散列因子和用户业务数据流建立连接时的散列因子不同，用户数据流恰好走故障链路而影响业务，而监控流经过的链路则有可能是正常路径。

端到端网络质量监控方案通过改变监控探针 IP 地址和监控目的 IP 地址，即改变散列算法中散列因子的方式，使监控数据流遍历链路，及时掌握网络拓扑变化，快速定位网络故障，并且通过分析监控数据，动态调整网络性能基线，收敛告警。此方案解决了传统网络质量监控系统遇到的故障发现不及时、故障定位不准确、无法掌握网络拓扑、性能基线设置单一等问题，可使日常运维工作变得更简单、高效。与传统网络质量监测方式比较见表 3。

5 结束语

根据散列算法用足够多的源目 IP 地址组合实现监控路径遍历性，通过单向监测方式绘制监测拓扑发现问题，用双向监测进一步定位故障，使用监控数据制定网络性能基线，收敛告警数据，建立跨域端到端网络质量监测系统可弥补传统网络质量监控手段的不足。后续可以考虑结合多种不同协议配合 ICMP 做路径监控，如使用 HTTP（hypertext transfer protocol，超文本传输协议）、UDP（user datagram protocol，用户数据报协议）。网络路径中间节点设备会对 HTTP 和 UDP 做转发策略，而针对 ICMP 分组，一些路由器、交换机等设备会为了不影响转发业务数据性能，对 ping 和 trace 协议分组做限制。根据 ICMP、HTTP、UDP 3 种协议各自优点，在不同场景结合使用的话可能会得到更加准确可靠的监测结果。

参考文献：

[1] 樊自甫, 伍春玲, 王金红. 基于 SDN 架构的数据中心网络路由算法需求分析[J]. 电信科学, 2015, 31(2): 42-51.
FAN Z F, WU C L, WANG J H. Requirements analysis of data center network routing algorithm based on SDN architecture [J]. Telecommunications Science, 2015, 31(2): 42-51.

[2] 谢海华. 有线 IP 城域网络可视化质量监控系统的建设与运用研究[J]. 无线互联科技, 2017(23).
XIE H H. Study on construction and application of visual quality monitoring system for CATV's IP metropolitan area network[J]. Wuxian Hulian Keji, 2017(23).

[3] 夏刚. 互联网环境的网络质量监测体系研究与实践[J]. 中国金融电脑, 2016(7): 48-51.
XIA G. Research and practice of network quality monitoring system in internet environment[J]. China Financial Computer, 2016(7): 48-51.

[4] 覃佐曼. 基于 SDN 的数据中心网络多路径负载均衡的研究[D]. 大连: 大连海事大学, 2017.
QIN Z M. Research on multipath load balancing in data center network based on SDN[D]. Dalian: Dalian Maritime University, 2017.

[5] 安禄. 基于等价多路径的数据中心网络流量优化问题研究[D]. 重庆: 重庆大学, 2014.
AN L. The optimization research of traffic engineering for data

表 3 与传统网络质量监测方式对比

	传统网络质量监控系统	端到端网络质量监控系统
监测方式	单向监测	单向、双向监测相结合
监测遍历性	单向监测且单一源目的地址无法实现监测链路遍历性	单向、双向监测方式相结合，且支持多个源地址和目的地址组合，可以实现监测链路遍历性
定位故障	单向监测方式无法判断是去程方向路径异常还是回程路径出现问题，无法定位故障	双向监测方式可以做到准确定位故障

center networks based on ECMP[D]. Chongqing: Chongqing University, 2014.

[6] 程光, 龚俭, 丁伟, 等. 面向 IP 流测量的散列算法研究[J]. 软件学报, 2005, 16(5): 652-658.
CHENG G, GONG J, DING W,et al. A hash algorithm for IP flow measurement[J]. Journal of Software, 2005, 16(5): 652-658.

[作者简介]

颜永明（1978−），男，中国电信股份有限公司上海分公司信息网络部综合运营监控中心副经理、高级工程师，主要研究方向为数据网络、云组网等。

陈兵（1970−），男，上海市信息网络有限公司总经理助理、高级工程师，主要研究方向为数据通信、大数据挖掘等。

许文杰（1986−），男，中国电信股份有限公司上海分公司信息网络部综合运营监控中心技术工程师、助理工程师，主要研究方向为数据网络。

电网企业财务健康诊断知识推理技术

万齐鸣[1]，王英军[1]，李有华[2]
（1. 北京中电普华信息技术有限公司，北京 100192;
2. 国网能源研究院有限公司，北京 102209）

摘　要：研究了电网企业财务健康诊断知识推理技术，将专家系统与神经网络相结合，采用专家系统框架，利用神经网络完成部分知识的存储与表示，克服了传统诊断方法的局限性，将领域专家的知识与神经网络的自学习能力有机融合在一起，实现对电网企业财务更健康、更准确的实时状态检测与智能诊断。

关键词：财务健康；检测诊断；知识推理；专家系统；神经网络
中图分类号：TP391.5
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018148

Knowledge inference techniques for financial health diagnosis of power grid enterprises

WAN Qiming[1], WANG Yingjun[1], LI Youhua[2]
1. Beijing China-Power Information Technology Co., Ltd., Beijing 100192, China
2. State Grid Energy Research Institute Co., Ltd., Beijing 102209, China

Abstract: The knowledge inference techniques for financial health diagnosis of power grid enterprises was studied. Combining experts system with neural network, expert system framework was used, finishing storage and representation of partial knowledge with neural network, thus overcoming the limitations of traditional diagnosis methods, organically integrating experts' knowledge in the field with self-learning abilities of neutral network. More accurate real-time status detection and intelligent diagnosis for financial health of power grid enterprise were realized.

Key words: financial health, detection and diagnosis, knowledge inference, expert system, neural network

1　引言

国家电网公司业务规模的扩大化、经营地点的分散化、组织人员的多元化、信息沟通的复杂化等，导致企业决策难度增加。国家电网公司对决策支持所需数据的准确性和时效性都有非常强烈的需求，要求财务部门更加快速地为企业决策提供更充分、及时、准确的信息以及适当的决策方法和决策建议，要求财务部门将更多的精力和资源投向战略和业务支持。

收稿日期：2018−01−23；修回日期：2018−04−10
基金项目：国家公司科技基金资助项目“电网集团公司管理会计关键技术及财务健康状态检测诊断研究”（No.5202011600U9）
Foundation Item: National Company Science and Technology “Research on Key Technologies and Financial Health Status Detection and Diagnosis of the Management Accounting of Power Grid Group Corporation” (No.5202011600U9)

电网企业财务健康诊断针对新形势下的财务管理变革和业财一体化决策需求，基于财务数据集中的业务背景，应用大数据和智能计算技术，以财务为视角，基于企业真实的业务财务数据流转，对于企业财务健康状态进行实时检测、计算和诊断，为企业业务科学决策和企业健康运行提供支持。

电网企业财务健康诊断可以有效帮助企业管理层及时了解企业财务的质量，有利于企业及其财务进行实时监控，实时把握企业生产、经营、管理现状，及时发现问题并采取处置措施，从而实现企业财务的健康运行。

2　电网企业财务健康诊断模型

电网企业财务健康诊断基于企业当前财务状态，利用智能计算技术，实时计算当前企业财务健康指数，对于异常指标进行对标分析和异常预警，利用知识推理技术，给出问题诊断及处置方案建议，并利用大数据分析挖掘技术，对于发现的问题进行问题成因分析，利用机器学习技术，对于诊断结果进行跟踪验证，并不断进行模型优化和知识库进化，从而达到检测诊断企业财务健康状态、实现业务优化提升的目的。

电网企业财务健康诊断模型如图 1 所示。

3　电网企业财务健康诊断技术的比较与选择

电网企业财务健康诊断可以参考和借鉴故障诊断技术。由于故障诊断原理不同，故障诊断方法也有不同。基于解析模型和基于信号处理的方法，由于知识表达能力的局限性，一般适用于故障监测和简单场合的故障诊断。而基于知识的方法，除了依赖数学模型以外，还具有较为丰富和灵活的知识表达能力以及问题求解能力，除进行离线诊断外，还可用于在线的故障诊断及故障处理。因此，采取基于知识的方法进行此项研究。

基于知识的方法主要包括灰色诊断方法、模糊诊断方法、神经网络诊断方法、专家系统诊断方法以及信息融合方法。其中，信息融合方法就是将多种方法融合在一起，包括专家系统与神经网络的融合、模糊系统与神经网络的融合以及遗传算法与神经网络的融合。

由于专家系统依赖于知识获取，在现实中存在知识获取困难、控制策略不灵活等缺点，难以保证实时诊断的实现。而人工神经网络采用神经元及其之间的连接权重来隐含处理问题的知识，能够处理复杂问题。此外，由于神经网络具有自学习能力，在故障诊断领域被广泛地应用。但神经网络诊断方法在训练样本的获取上存在一定的

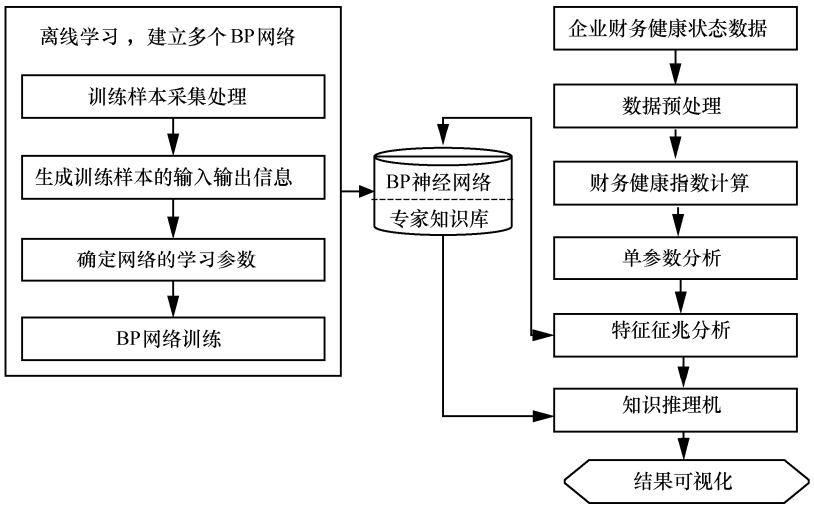


图 1　电网企业财务健康诊断模型

困难，此外，神经网络方法往往忽视了相关专家多年的经验积累，网络权值的表达也不易理解。专家系统适合逻辑，神经网络长于思维，将它们进行有效的集成能起到优势互补的作用，使得建立的检测诊断系统同时具有很强的学习能力、解释能力和推理能力。

电网企业财务健康诊断模型将专家系统与神经网络相结合，利用专家系统作为框架，在其知识的存储和表示上，将神经网络融入其中。在进行财务健康检测诊断时，利用信息融合方法建立一个智能计算诊断系统，使神经网络与专家系统相辅相成。通过构建智能计算诊断知识库、神经网络知识学习以及专家系统推理，完成对企业财务健康状态的检测与诊断。

4 基于专家系统与知识推理的财务健康诊断

专家系统是一个具有大量专门知识与经验的程序系统，它应用人工智能和计算机技术，根据某领域一个或多个专家提供的知识和经验进行推理和判断，模拟人类专家的决策过程，以便解决那些需要人类专家处理的复杂问题。专家系统通常由人机交互界面、知识库、推理机、解释器、综合数据库、知识获取 6 个部分构成。以下重点从知识库构建、知识表示、知识推理机阐述电网企业财务健康诊断的知识推理机制。

4.1 知识库的构建

由于电网企业财务健康状态的多样性和异常状态表示的差异性，专家系统的知识库采用多层级模式，分类建立数据库、事实库、规则库。数据库用于存储电网企业财务状态，包括历史数据和当前状态数据；事实库存储专家知识与专家经验；规则库包括诊断规则库和元知识规则库。诊断规则库又包括用于异常状态诊断的知识、用于异常状态原因分析的知识和用于消除异常状态的处置措施的知识。

专家系统知识库构建的层级模型如图2所示。

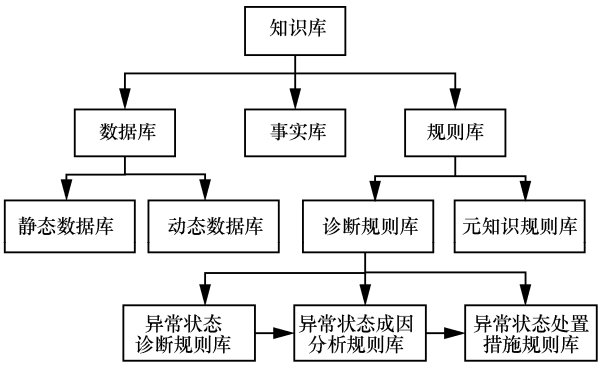


图 2 专家系统知识库层级模型

4.2 知识的表示

专家系统的知识表示主要采用谓词结构、产生式规则两种表达方式，根据知识的特点辅以框架表示方法。由于专家的经验知识主观因素较大，因此以其为依据产生的规则并非完全可信。为了得到更加客观的结果，可在产生式规则中增加可信度因子。

（1）财务健康状态指标数据用谓词逻辑表示

在谓词逻辑中，事实由一个关系和一些有关系的个体组成。用 $P(x_1,x_2,\ldots,x_n)$表示 n 元谓词计算式，其中，P 为 n 元谓词，$(x_1,x_2,\ldots,x_n)$为客体变量或变元。

假设有命题：企业财务状况=average（一般），要求短期偿债能力=good（好）且长期偿债能力=average且经营能力=good且获利能力=average。

用 $I(x)$表示“企业财务状况=average”，$L(x)$表示“短期偿债能力=good”，$H(x)$表示“长期偿债能力=average”，$P(x)$表示“经营能力=good”，$Q(x)$表示“获利能力=average”。上述命题用谓词计算式可以表示为：

$$(x)(I(x))\rightarrow(L(x)\wedge H(x)\wedge P(x)\wedge Q(x)) \quad (1)$$

（2）检测型的知识采用产生式规则表示

基于规则的产生式系统是一种很适合于表达因果关系的表示模式，是目前专家系统中最为普遍的一种表达方法。产生式的规则：其“if-then”结构接近人类思维和会话的自然形式，易于人们

在特定情况下的行为知识的表达和编码。企业财务健康检测诊断的检测型知识可以采用产生式规则表示。实际情况中，如果企业状态检测知识比较模糊，可增加可信度因子 CF 表征这些事实的经验可信度值。

在产生式系统中，论域的知识分为两种：一是事实知识，另外一种是用产生式规则表示的推理性知识。产生式规则是一个“如果满足某条件，就采取某些操作”的语句。即：

if　条件/条件组

then　结果。

假设“短期偿债能力=unsatisfactory（差）”的原因集合：

F 短期偿债能力差={（流动比率=unsatisfactory，0.5），（存货周转率=acceptable（尚可），0.2），（应收账款周转率=unsatisfactory，0.3）}。

可以用产生式规则表示为：

if（流动比率=unsatisfactory，0.5）and（存货周转率=acceptable，0.2）and（应收账款周转率=unsatisfactory，0.3）

then 短期偿债能力=unsatisfactory。

4.3　知识推理机

专家系统的知识推理过程是通过推理机完成的，推理机就是智能系统中用来实现推理的程序，是知识系统的主要部分。推理机的基本任务就是在一定控制策略指导下，搜索知识库中可用的知识，与数据库匹配，产生或论证新的事实。搜索和匹配是推理机的两大基本任务。

专家系统的知识推理包括两个基本问题：一是推理方法，二是推理的控制策略。推理方法研究的是前提与结论之间的种种逻辑关系及其信度传递规律等，控制策略是推理机的核心部分，它的主要任务是解决知识的选择与应用的顺序，也就是确定搜索方式和搜索方法，目的是限制和缩小搜索的空间，使原来的指数型困难问题在多项式时间内求解。从问题求解角度来看，控制策略亦称为求解策略，它包括推理策略和搜索策略两大类。

作为专家系统的组织控制机构，推理机能通过运用由用户提供的初始数据，从知识库中选取相关的知识，并按照一定的推理策略进行推理，直到得出相应的结论。在设计推理机时应考虑推理方法、推理方向和搜索策略。

（1）推理方法

在现实中，事物的特征并不总是表现出明显的是与非，同时还可能存在着其他原因，如概念模糊、知识本身存在着可信度问题等。专家系统的问题求解不像经典数学、物理等学科那样具有严密性和精确性，领域专家的知识和人们要处理的信息往往是不确定的。因此为了把这些不确定的知识表示在专家系统中，并且能用这些形式化、不确定的知识进行判断、推理和决策，在专家系统中往往要使用不精确推理方法。

在不精确推理模型中，财务健康检测诊断使用模糊推理，其理论基础是模糊集理论以及在此基础上发展起来的模糊逻辑。

（2）推理方向

推理方向有 3 种：正向推理、反向推理、正反向混合推理。

正向推理，即从已知的事实出发，向结论方向进行推导，直到推出正确的结论。它的大体过程是：系统根据用户提供的原始信息与知识库中规则的前提条件进行匹配，若匹配成功，则将该知识块的结论部分作为中间结果，利用这个中间结果继续与知识库中的规则进行匹配，直到得出最后的结论。

如果说正向推理是自下而上的方式，那么反向推理与之相反。反向推理是以自上而下的方式，以反向验证的方式进行推理。反向推理从目标出发，沿着推理路径追溯到事实。与正向推理相比，反向推理具有很强的目的性。

正反向混合推理是指先根据给定的不充分的

原始数据或证据向前推理，得出可能成立的诊断结论，然后，以这些结论为假设，进行反向推理，寻找支持这些假设的事实或证据。

企业财务健康检测诊断平台在检测到数据异常状态时，先采取正向推理的方法根据目前异常状态（现象）初步推理出可能的原因，然后通过反向验证（反向推理）进一步求证目标，排除多个原因中的一部分内容，然后再进行正向推理。采用正、反向推理相结合的混合方式，既可以避免正向推理的盲目性，也可避免反向推理中初始目标选择的盲目性，是一种有益的互补方式。

（3）搜索策略

知识推理机的搜索策略中，宽度优先搜索、深度优先搜索两种算法的应用非常广泛。

宽度优先搜索方法按照一层一层的步骤搜索，即首先搜索与起始节点直接相连的下一层节点，再搜索所有与下层节点直接相连的更下层节点，依次类推。

与宽度优先搜索的机制不同，在深度优先搜索中，深度越大的节点越先得到扩展。深度优先搜索方法主要存在两个缺点：有可能出现无穷递归的情况，从而搜索不到需要的解；即使搜索到也可能不是最短路径。

对于电网企业财务健康检测知识推理，在知识库的规则查找路径方式上，正向推理可采取深度优先搜索策略，以找到问题原因，而反向推理可采用宽度优先策略。

4.4 实证研究

假设电网企业财务健康所有要检测的对象可能发生的状态组成的空间为 S，检测过程中可测量数据的参数特征组成的空间为 Q，某个状态 s 与其对应的特征 q 之间的关系用映射 g 表示，g：$S\rightarrow Q$。反之，某一特征 q 也对应于企业财务健康的确定状态 s，即存在映射 f，f：$Q\rightarrow S$。电网企业财务健康诊断的目的是根据测量到的特征向量判断当前企业财务所处的状态，即映射 f。

假设与参数有关的企业财务健康可能发生状态是有限的，为 n，正常的状态为 s_0，财务指标有 m 个，则状态空间可以表示为 $S=\{s_0,s_1,\ldots,s_n\}$。企业财务健康状态 s_i 对应的参数特征值为 $Q_i=\{q_{[i,1]},q_{[i,2]},\ldots,q_{[i,m]}\}$。企业财务健康状态的检测通常是由实时数据、历史数据联合进行判断完成，因此特征参数的取值大多数情况下为集合$\{Q_i\}$。

当企业财务状态检测不健康时，首先依据当前的异常状态利用专家系统进行正向推理，得到问题发生的可能原因，再利用反向推理的方法进行验证，排除一部分可能的原因后，再进行正向推理。通过正向推理和反向推理二者结合的方法，避免了两种推理过程中的盲目性，得到最佳推理诊断结果。

电网企业财务健康状态的检测诊断，实际上就是根据得到的 i 个特征参数对状态进行分类，或称为对状态进行模式识别。这个过程包括 3 个步骤：检测企业财务监控指标状态值；异常特征或征兆提取；由异常特征及其他知识确定企业财务健康的状态，完成异常诊断和预警。其原理如图 3 所示。

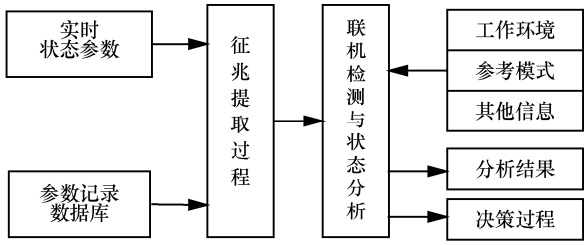


图 3　异常状态监测诊断预警原理

5　基于神经网络与知识学习的模型优化

在电网企业财务健康诊断模型中，核心知识推理机用到的诊断规则库可以考虑通过 BP（back propagation）神经网络进行知识的存储与知识表示。

由于企业财务健康异常状态的多样性、复杂性，包括异常状态特征、异常状态成因、异常状

态处置措施等相关的信息，依据异常状态层级、异常状态类别划分，可以建立多个神经网络。

以图 4 中的典型异常状态与特征征兆之间的知识图为例，说明 1 个子神经网络的建立过程。假设成本费用异常模块包括购电成本异常、科技研发成本异常、基建维修成本异常 3 个子网络。购电成本异常子网络包括 4 个输入神经元（火电、风电、水电、太阳能发电），1 个输出神经元，由于输入、输出层神经元较少，可以采用三层神经网络，隐含层神经元可以取为 3 个，其拓扑结构如图 4 所示。

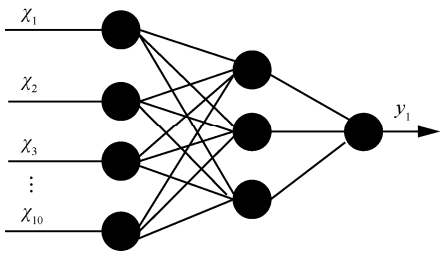


图 4　神经网络拓扑结构

神经网络不仅可以用于知识存储与知识表示，其结构特征也可作为知识推理机的推理规则，进行异常检测、诊断，并给出处置建议。

神经网络同时还具有自学习能力，可以对于模型进行优化提升。神经网络的学习属于 Delta 算法，是一种有监督的学习算法，根据网络的实际输出与期望输出之间的差别调整各个连接权值，直至误差在可接受的取值范围之内。其学习算法可以采用改进的最小梯度法，改善标准学习算法中局部极小、收敛速度慢的缺点。神经网络算法要求激活函数处处可导，另外学习速率根据误差函数序列的增益情况进行适应性调整，减少迭代次数、避免震荡甚至发散情况的发生。

通过神经网络与机器学习，结合财务健康实际的检测、诊断、处置、效果跟踪的实际数据作为样本，可以不断完善知识库的知识表示与知识推理规则，使其更接近于企业实际的应用场景，从而使得检测诊断结果更为准确，处置建议更为恰当，可以更好地发挥企业财务健康诊断的应用效果。

6　结束语

电网企业财务健康诊断基于企业当前生产、经营、管理状态，利用智能计算技术，实时计算当前企业财务健康指数，利用知识推理技术，给出诊断解决方案，从而达到监控企业财务健康状态、实现规避企业财务风险的目的。

电网企业财务健康诊断理论研究成果目前正由北京中电普华信息技术有限公司进行成果转化，设计和开发电网企业财务健康检测诊断系统，并结合国网信息通信产业集团公司实际案例，构建企业财务健康监控指标体系，设计财务健康指数，进行数据采集和数据清洗、模型构建，以满足企业财务健康检测诊断的应用需求。实际应用效果由于系统正在建设中，尚有待验证。

本文研究电网企业财务健康诊断知识推理技术，将专家系统与神经网络相结合，采用专家系统框架，利用神经网络完成部分知识的存储与表示，克服了传统检测诊断方法的局限性，将领域专家知识与神经网络的自学习能力有机融合在一起，实现对电网企业财务更健康、更准确的实时状态检测与智能诊断。

参考文献：

[1] 梁烂英. 企业财务健康评价指标体系研究-来自信息技术类上市公司的实证分析[D]. 杭州: 浙江大学, 2011.
LIANG L Y. Research on index system of corporate financial health evaluation-empirical analysis from information technology listed companies[D]. Hangzhou: Zhejiang University, 2011.

[2] 曹旭, 曹瑞彤. 基于大数据分析的网络异常检测方法[J]. 电信科学, 2014, 30(6): 152-156.
CAO X, CAO R T. Network anomaly prediction method based on big data [J]. Telecommunications Science, 2014, 30(6): 152-156.

[3] 杨慕涵. 电力行业上市公司财务预警研究[D]. 西安: 西安工业大学, 2013.
YANG M H. Research on financial early-warning of listed companies in electric power industry[D]. Xi'an: Xi'an Techno-

logical University, 2013.

[4] 李传煌, 孙正君, 袁小雍, 等. 基于深度学习的实时DDoS攻击检测[J]. 电信科学, 2017, 33(7): 53-65.
LI C H, SUN Z J, YUAN X Y, et al. Real-time DDoS attack detection based on deep learning[J]. Telecommunications Science, 2017, 33(7): 53-65.

[5] 任庆霜, 司景萍. 基于神经网络的汽车电喷发动机故障诊断[J]. 内燃机与配件, 2010(4): 33-35.
REN Q S, SI J P. Fault diagnosis of automotive EFI engine based on neural network[J]. Internal Combustion Engine & Parts, 2010(4): 33-35.

[6] 姜红红, 张涛, 赵新建, 等. 基于大数据的电力信息网络流量异常检测机制[J]. 电信科学, 2017, 33(3): 134-141.
JIANG H H, ZHANG T, ZHAO X J, et al. A big data based flow anomaly detection mechanism of electric power information network[J]. Telecommunications Science, 2017, 33(3): 134-141.

[7] 李长河, 曹广传.一种通用型模糊环境智能故障诊断系统的实现[J]. 西安理工大学学报, 2006, 22(3): 323-326.
LI C H, CAO G C. Implementation of a general fault diagnosis system based on fuzzy ambient intelligence[J]. Journal of Xi'an University of Technology, 2006, 22(3): 323-326.

[8] 张国伟, 徐宏, 毛红奎. 铝铸件缺陷专家系统知识推理策略研究[J]. 铸造技术, 2011, 32(12): 1690-1693.
ZHANG G W, XU H, MAO H K. Study on expert system inference analysis control strategy of casting defects in aluminum castings[J]. Foundry Technology, 2011, 32(12): 1690-1693.

[9] 章伟聪. 基于神经网络的汽油发动机的故障诊断专家系统初探[J]. 浙江万里学院学报, 2005, 18(2): 27-31.
ZHANG W C. Fault diagnosis expert system of auto engines based on artificial neural networks and expert system [J]. Journal of Zhejiang Wanli University, 2005, 18(2): 27-31.

[作者简介]

万齐鸣（1971−），男，北京中电普华信息技术有限公司发展研究中心高级技术专家，主要研究方向为大数据、智能计算、知识管理、搜索引擎。

王英军（1974−），男，北京中电普华信息技术有限公司发展研究中心主任、工程师，主要研究方向为企业信息化咨询、财务信息化、项目管理。

李有华（1974−），男，国网能源研究院有限公司财务与审计研究室主任、高级审计师，主要研究方向为财务与审计。