面向业务感知的算网融合关键技术研究

Service-Aware Computing and Network Convergence Technology

周吉喆/ZHOU Jizhe，杨思远/YANG Siyuan，
王志勤/WANG Zhiqin

（中国信息通信研究院，中国 北京 100191）
(China Academy of Information and Communications Technology, Beijing 100191, China)

DOI：10.12142/ZTETJ.202205002
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.TN.20221012.1747.002.html
网络出版日期：2022-10-13
收稿日期：2022-08-14

摘要：为解决现有算网融合方案存在的完整业务建模缺失、算网资源与业务需求失配、系统多性能指标折中关系不清的问题，提出了基于服务化架构的算网融合关键技术。通过建模微服务与业务性能相关性业务模型，设计适配业务需求的算网融合资源调度方案，并研究面向系统多种性能联合优化的算网融合一体化编排策略，满足新一代移动通信网络建设中业务高性能、系统高效能、算网深融合的综合发展需求。

关键词：算网融合；服务化架构；资源管理

Abstract: The current schemes of computing and network convergence (CNC) resource allocation face three challenges: the lack of a general profile and model of service, the mismatching between computing-network resource allocation and service needs, and the missing analysis of the tradeoff relation among various system key performance indicators (KPIs). To solve these problems, the combination between service-based architecture and CNC is studied, and the service-aware CNC technology is introduced. By leveraging the orchestration strategy of computing and network for the system-level optimization on multiple KPIs, the service-aware CNC can greatly enhance the performance of applications, improve the efficiency of the system and strengthen the convergence of computing and network.

Keywords: computing and network convergence; service-based architecture; resource management

1 算网融合发展现状

随着新一代科技革命和产业革命的持续演进，以工业互联网、车联网、虚拟及增强现实和数字孪生为代表的数字化、网络化、智能化业务蓬勃发展。据IMT-2030（6G）推进组《6G典型场景和关键能力白皮书》[1]预测：到2030年，各类移动终端连接数将超过184亿，月均流量将高达5 835 GB。

新数字业务的急速增长将需要更强的网络数据处理与计算能力。当前，仅依赖云计算或单一边缘计算的端到端数据处理与计算方式存在业务多元需求但算力供给不均、业务快速变化但算力延迟响应、业务分布泛化但算力调配集中等系统性矛盾，难以有效支撑业务发展。同时，随着手机、智能汽车等各类移动终端设备计算能力的提升，云计算、边缘计算等网络算力呈现多样化发展的趋势。根据《中国算力发展指数白皮书》[2]，2020年中国算力规模达到135 EFlops（亿次浮点运算），年增长率达到55%，算力规模持续增加。算力规模的快速增长与算力的多样性发展将驱动网络算力分配向泛在化演进，即网络内部分布着不同计算能力的算力节点。为此，如何高效调用分布式算力，有效匹配并服务多样泛在的算力资源，从而实现算力与网络的深度融合，是新一代移动通信网络发展的重要问题。

面向上述业务需求，中国高度重视算力与网络融合（以下简称“算网融合”）技术研究，将其列为6G网络技术储备的重要内容[3-6]。2021年5月，国家发展和改革委员会等4部门联合印发的《全国一体化大数据中心协同创新体系算力枢纽实施方案》指出，需要加快建设全国算力一体化的新型算力网络架构，进行算网资源的优化配置，实施“东数西算”工程[7]。在标准化方面，全球的标准化组织积极力推动算网融合领域的标准进程。国际电信联盟（ITU）已将算力网络架构与算网融合需求相关的标准做了立项，探索算网融合逻辑功能与信息交互机制。欧洲电信标准化协会（ETSI）聚焦动态路由、分流等算网融合技术研究，开展通信与计算融合的高效网络系统研究。互联网研究任务组（IRTF）成立在网计算研究工作组，面向可编程网络设备研究“计算+

基金项目：国家重点研发计划（2021YFB2900200）；中国博士后科学基金（2022M713475）

转发”的算网融合网络内生功能。中国标准化组织也积极推动算网融合、算力网络相关的研究和立项。IMT-2030（6G）推进组成立算力网络研究组，旨在推动业界达成算网融合技术研究共识，加快关键技术攻关与系统设计进程。

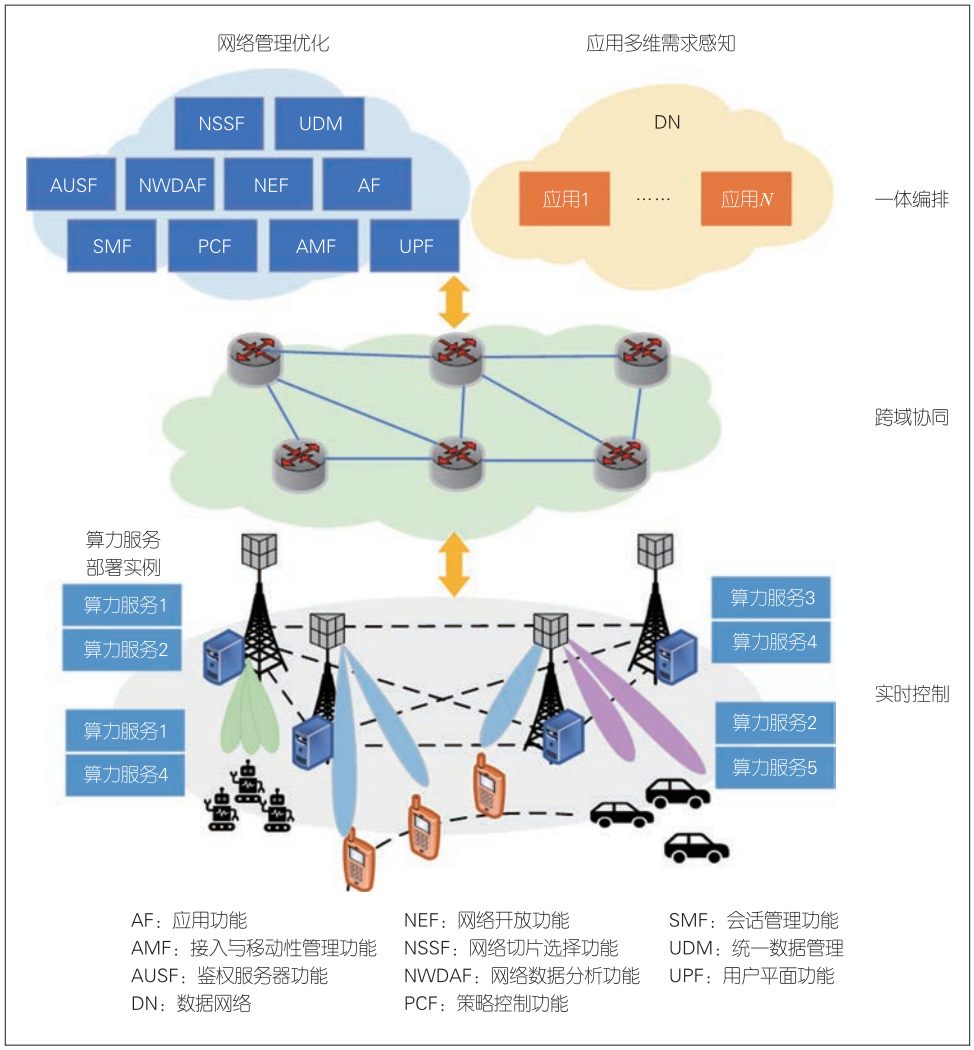
目前，业界主要聚焦基于互联网协议（IP）技术的算力感知与算网融合技术研究，通过感知业务服务质量（QoS）要求与分布式算网资源状态，将业务调度到合适的算力节点。但目前这些研究仍然面临分布式算力协同机制匮乏、算网资源管理灵活性与动态性受限、业务性能与可靠性保障机制缺乏等挑战。针对未来泛在移动业务接入与极致性能需求，算网融合技术发展将与未来移动通信网络服务化、分布式的演进方向紧密结合，进一步细化算力度量与感知颗粒度，建立网络泛在算力资源与实时网络状态的弹性协同管理机制，为业务性能指标提供确定性保障。

2 基于服务化架构的算网融合技术演进趋势

5G网络确定了基于服务化的网络架构，并利用虚拟化技术灵活部署核心网控制面网元，建立网元间的统一接口与交互机制。未来网络将进一步深化端到端网络的服务化进程[7-8]。云原生技术能够将网络功能拆解为松耦合的微服务，并利用容器等虚拟化技术将微服务独立运行并按需部署，以支持更加广泛的应用场景与灵活高效的网络服务能力，拓展网络服务化架构应用边界。

结合未来网络全面服务化的演进趋势，基于服务化架构的算网融合资源管理能够在算力协同与业务性能的角度适配分布式算力高效调度的技术路线，增加算网资源管理的灵活性，助力自配置、自优化的网络功能编排升级。同时，面向微服务的业务治理方法能够通过按需灵活的算网资源进行管理，满足实时、跨域、智能化等多样化的业务场景与性能要求。在算力协同方面，与传统面向业务整体的资源分配方案不同，运行微服务所需的算网资源规模相对较小，能够充分调用算力规模较小的算力节点，并通过多点算力协作提高算网资源利用率与系统容量。因此，基于服务化架构的服务编排思路与分布式算力高效调用的技术路线相匹配。在业务性能方面，以微服务为对象的资源分配方案将算网资源进行统一封装，增加了算网服务的多样性与定制性。同时，基于微服务的算力服务编排允许每个微服务单独部署在不同的容器或算力节点上。这使得同一微服务可以被不同业务所调用，更加灵活地通过微服务冗余部署实现负载均衡。因此，基于微服务的算网资源管理增加了微服务调用的灵活性，进一步提升了时延、可靠性等业务性能指标，增强了多业务并运行下的网络服务能力。面向业务感知的服务化架构算网融合技术愿景如图1所示。

为最大程度发挥服务化架构在适应性、弹性、可靠性上的优势，算网融合通过对算力与网络等多维资源进行统一编



▲图1 面向业务感知的服务化架构算网融合技术愿景

排与调度，以算力服务的形式向业务提供按需调度、灵活适配的网络服务，拓宽了未来网络的新价值。

3 基于服务化架构的算网融合研究挑战

为了充分发挥微服务等服务化相关技术在算网资源管理问题上的赋能效用，我们需要着力解决以下两方面的问题。首先，需要对拆解的微服务与业务整体的相关性进行合理的分析与数学建模。由于微服务间存在功能上的依赖关系，微服务单体性能与业务整体性能间存在制约关系，因此，基于微服务的科学业务建模方法是算网资源高效利用的研究基础。其次，需要面向业务需求感知与业务特征，建立算网资源的协同控制机理。在面向微服务设计算网资源分配方案时，由于微服务的算网资源需求不同，且业务的微服务逻辑结构复杂，需要充分分析微服务单体与微服务局部交互对算网资源分配的影响。

3.1 基于微服务的算网融合资源管理研究现状

目前，学术界主要聚焦无线接入网、边缘计算等特定场景下的基于微服务的算网资源管理研究。为了实现算力与网络资源的高效调度并最终体系化地建立基于微服务的业务模型，应首先充分分析与提取微服务的需求、特征与相关性。

现有普遍使用的基于微服务的业务建模方法主要有两种。第一种是将业务建模成有序的微服务集合，微服务间存在线性依赖且在时间上有前后逻辑关系[9–13]。例如，人脸表情识别微服务需要在人脸识别微服务后运行。第二种方法是将业务建模成有向无环图（DAG），只有当有向边上的前导微服务完成后才能进行其指向的后续微服务[14–18]。在这种情况下，微服务存在不止一个前导微服务，其部署方案与性能受到所有前导微服务集合的影响。这两种业务建模方法均考虑了实际业务运行过程中业务模块间存在的逻辑关系。由于业务微服务间的逻辑关系，前导微服务的部署方案与性能会影响后续微服务的通信资源分配与算力节点部署决策。因此，需要根据微服务间的关联性设计算网资源分配方案，以提高业务性能。基于DAG的业务建模方法在一定程度上考虑了业务微服务间的复杂逻辑关系，并设计了相应算网资源分配方案，提高了业务性能。

3.2 基于服务化架构的算网融合面临的挑战

现有工作虽然能够面向微服务进行算网融合的资源协同分配，但仍存在微服务与业务性能相关性的业务建模方法缺失、算网融合资源调度与业务需求失配的问题，未能建立系统融合性能最优的算网融合编排管理机制。基于服务化架构的算网融合研究面临以下一些挑战。

（1）微服务与业务整体相关性建模缺失：微服务间存在逻辑关系与交互行为，因此微服务单体部署决策影响业务整体服务水平。现有工作未充分考虑不同微服务的异质化特性，仅建模了微服务间单向的逻辑关系，未对多个微服务间信息交互与相互依赖的逻辑关系进行建模与量化，无法真实反应业务需求与微服务逻辑关联的结构特征。因此，目前缺少通用的基于微服务的业务建模方法。

（2）算网融合资源调度与业务需求失配：为了最大程度地提升业务性能，需要结合微服务的资源需求与相关性特征，设计对应的算网资源分配方案。现有基于微服务的算网融合资源管理方案并未完整分析微服务间交互行为对业务性能的影响，这导致资源分配无法满足实际业务需求。

（3）系统多种性能指标折中关系不清晰：在多业务场景下，需要在保证业务QoS的前提下，最大程度提升网络整体性能与容量。现有算网融合资源管理方案通常以业务时延、吞吐量等单一性能作为优化目标，对可靠性、系统容量、算网资源开销等性能指标间折中关系的研究并不清晰。因此，现有算网融合资源管理方案无法实现系统整体服务能力最优的目标。

4 基于服务化架构的算网融合关键技术

为确保基于服务化架构的算网资源的灵活高效调度，业界亟需解决业务建模缺失、算网资源与业务需求失配、系统多性能权衡不清等关键问题，突破基于云原生的算力服务与业务建模、业务需求适配的算网融合资源调度、面向系统整体性能优化的算网融合编排等技术瓶颈。基于服务化架构的算网融合关键技术研究将推动业务性能与网络整体效率的优化。构建普适的业务模型可以为适配业务需求的最优算网融合资源调度提供模型基础。对业务模型与分布式算网资源拓扑的映射关系与资源分配方法的研究，能够为复杂环境下的系统整体性能优化提供理论基础，从而指导系统整体服务能力最优的算网融合服务编排。

举例来说，基于服务化架构的算网融合技术能够有效赋能增强现实类应用场景。通过对使用者看到的现实世界进行数字化环境渲染与增强，增强现实技术实现了使用者在虚实世界的交互体验。目前增强现实技术已广泛应用于智慧工厂中的设备巡检、智慧教育中的教学互动等多种场景。增强现实类应用通常包含视频采集模块、跟踪模块、映射模块、物体识别模块、渲染模块等功能模块。各个模块的开发与部署可以相互独立，且模块间存在复杂的逻辑关系。例如，跟踪模块需要根据视频采集模块采集的图片与当前渲染模块的渲

染结果进行特征点学习，完成对镜头与目标物体的跟踪同步。同时，映射模块根据跟踪结果更新虚拟世界模型，并反向为跟踪模块提供新的特征点参数。增强现实类应用对算力的要求较高，而大多数移动终端与头显设备受限于算力与能耗，无法完成高清视频渲染与实时虚实交互。基于服务化架构的算网融合技术能够充分调度云边端算力资源，满足业务性能要求并提升网络整体容量。首先，利用基于云原生的业务建模方法能够将不同模块视为微服务，从而量化其计算量、传输数据量等参数，并衡量不同模块间的复杂逻辑关系与交互行为。其次，根据微服务的算网需求与相互作用关系，进一步利用算网融合资源调度方法将增强现实类应用模块部署在最优的算力节点（如移动终端、边缘计算等）上，以满足实时交互应用需求。最后，面向多业务、动态等复杂网络环境，利用算网融合服务编排能够提高系统时延、可靠性、算网资源利用率等多种系统性能指标，从而显著提升系统整体效率。

4.1 基于云原生的算力服务与业务建模

云原生技术广泛地应用在复杂信息技术（IT）业务的开发、维护等环节，能够提高IT业务交付与后期管理效率。将云原生技术应用在算网融合中，能够解决现有算力服务单体庞大、部署僵化的问题，实现轻量级、定制化的算力服务管理。另外，云原生技术将底层网络、算力、存储等多维度资源封装为微服务，并独立地部署在算力节点上。这样能够更大程度地调度网络分布式异构的算力资源，实现算网资源协同控制。基于微服务的算力服务类型包括两种：第一种是移动通信网络的网络功能，例如核心网的切片管理、接入网的基带处理等；另外一种是网络内生支持的新型算力服务，例如人工智能（AI）训练等。组合不同的算力服务，能够满足多样化业务需求。

为了更好地提升业务性能，需要深入挖掘业务微服务间的逻辑关系和交互行为，建立通用的业务微服务模型，为算网融合资源协同调度管理提供模型基础。首先，对单体微服务的计算、通信等资源需求进行合理性分析，并建模量化微服务对业务整体性能的相关性水平，完善微服务单体模型参数。其次，分析微服务间的逻辑关系与交互行为，进而构建微服务间的逻辑关联结构。在传统业务建模方法的基础上，不仅需要考虑两个微服务间“单向线性”的逻辑关系，还要进一步建模两个或多个微服务间“环形交互”的复杂逻辑结构。由于前导微服务在资源需求与性能上的差异，不同前导微服务对后续微服务的部署方案与性能表征的影响力存在不同。因此，还需进一步量化逻辑关系上的前导微服务对后续微服务的影响力，建模微服务间的影响力参数。最后，根据微服务单体需求和逻辑关联结构，利用有向图的方法健全业务整体模型。有向图的顶点代表一个微服务，其特征参数包括计算能力要求、重要性水平等，有向边代表两个微服务间的逻辑关系，其特征参数包括微服务的业务传输量、影响力水平等。

4.2 业务需求适配的算网融合资源调度

基于微服务的业务建模将算力服务分布式部署在异构算力节点上。由于网络算力节点的计算资源有限，且节点间通信能力受到带宽、信道状态等通信资源与环境因素影响，因此需要根据业务需求与网络状态，对网络分布式的计算与通信资源进行整合与协同控制。研究建立业务模型与算网物理资源拓扑间的映射关系，能够对算力服务进行最优部署与灵活调度，最大程度提升业务性能。

为实现算网融合资源的最优调度，首先，分析算网物理资源拓扑与业务微服务模型的映射关系，构建面向业务需求与特征的算网资源优化问题。算网物理拓扑也可以利用图论的方法建模算力节点连接拓扑，并通过量化计算能力、通信能力等参数，构建算网资源模型。由于业务模型与算网资源模型具有不同的参数体系，可以通过在业务性能指标和算网资源维度上对应统一的方式，构建两者的关联模型，进一步研究算网融合资源分配优化问题。其次，对上述算网融合资源优化问题进行分解，通过降低优化变量间的关联性，降低问题求解的复杂度。该算网融合资源优化问题的复杂度主要来源于微服务间的逻辑关系与交互行为，因此，解决问题的关键在于对业务的有向图模型进行分解，形成多个算网融合资源优化的子问题。最后，根据分解后的微服务间逻辑关联性，建立子问题优化方向与整体方案设计目标的关联与统一，实现全局资源的最优配置。可以利用分布式学习、博弈论等方法，实现多智能体间的协同迭代学习，并最终获得最优的算网融合资源调度策略。

4.3 多目标联合优化的算网融合编排

随着网络规模与业务多样性的增加，算网融合资源优化目标不仅仅包括业务时延、吞吐等性能指标，还需要考虑业务可靠性、确定性及网络能效、算网资源利用率、算网管理开销等多种指标。因此，在适配业务需求的基础上，需要进一步考虑面向系统多性能联合优化的算网融合编排技术研究，通过对多指标权衡与协同管理，实现网络不确定性下的系统承载能力与效率提升。

首先，假设确定性的网络场景，并根据微服务的特征参

数，研究时延、可靠性、算网资源利用率等多种性能指标间的折中关系。多种性能指标的折中关系可以进一步建模成受约束的算网融合资源分配优化问题。通过对该优化问题进行求解，可以获得性能指标间相关性的数学表征，并能够给出算网融合编排指导。其次，面对网络不确定性场景，针对动态业务到达、算力节点中断等情况，利用概率论等方法研究灵活动态的算网融合编排策略。可以根据历史经验，利用梯度下降等方法设计线上算网融合编排策略，根据系统整体性能要求自适应地调整分布式算网资源分配。

5 结束语

本文探索了面向业务感知的服务化架构算网融合的资源管理技术研究，旨在推动云原生、微服务等技术在算网融合中的应用，引领通信与计算融合的新型网络设计范式。

参考文献

[1] IMT-2030(6G)推进组. 6G典型场景和关键能力白皮书 [R]. 2022

[2] 中国信息通信研究院. 中国算力发展指数白皮书 [EB/OL]. [2022-09-05]. http://www.caict.ac.cn/kxyj/qwfb/bps/202109/t20210918_390058.htm

[3] China Mobile, China Telecom, China Unicom, et al. 5G-advanced technology evolution from a network perspective (2021) [EB/OL].[2022-09-05]. https://www. huawei. com/en/events/5g-advanced-technology-evolution-network-perspective

[4] IMT-2030(6G)推进组. 6G网络架构愿景与关键技术展望白皮书 [R]. 2021

[5] PELTONEN E, BENNIS M, CAPOBIANCO M, et al. 6G white paper on edge intelligence [EB/OL]. (2020-04-30) [2022-09-07]. https://arxiv. org/abs/2004.14850

[6] WIKSTROM G, PERSSON P, PARKVALL S, et al. 6G-connecting a cyber-physical world [EB/OL]. [2022-09-06]. https://www. ericsson. com/en/reports-and-papers/white-papers/a-research-outlook-towards-6g

[7] 国家发展改革委, 中央网信办, 工业和信息化部, 等. 全国一体化大数据中心协同创新体系算力枢纽实施方案 [EB/OL]. [2022-09-06]. https://www.ndrc.gov.cn/xwdt/tzgg/202105/t20210526_1280839.html?code=&state=123

[8] IMT-2030(6G)推进组. 6G网络架构愿景与关键技术展望白皮书 [EB/OL]. [2022-09-06]. https://mp.weixin.qq.com/s/JXqJbVTvbPvJE_Tl5jRXKA

[9] 中国移动通信有限公司研究院. 6G服务化RAN白皮书 [EB/OL]. [2022-09-06]. http://cmri.chinamobile.com/insight/technology/7319.html

[10] WANG M, CHENG B, FENG W D, et al. An efficient service function chain placement algorithm in a MEC-NFV environment [C]//Proceedings of 2019 IEEE Global Communications Conference. IEEE, 2019: 1-6. DOI: 10.1109/GLOBECOM38437.2019.9013235

[11] CAO Z, ZHANG H G, CAO Y, et al. A deep reinforcement learning approach to multi-component job scheduling in edge computing [C]//Proceedings of 2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN). IEEE, 2019: 19-24. DOI: 10.1109/MSN48538.2019.00018

[12] CHEN L L, XU Y C, LU Z H, et al. IoT microservice deployment in edge-cloud hybrid environment using reinforcement learning [J]. IEEE Internet of Things journal, 2021, 8(16): 12610-12622. DOI: 10.1109/JIOT.2020.3014970

[13] WANG S Q, ZAFER M, LEUNG K K. Online placement of multi-component applications in edge computing environments [J]. IEEE access, 2017, 5: 2514-2533. DOI: 10.1109/ACCESS.2017.2665971

[14] CHEN J W, YANG Y J, WANG C Y, et al. Multitask offloading strategy optimization based on directed acyclic graphs for edge computing [J]. IEEE Internet of Things journal, 2022, 9(12): 9367-9378. DOI: 10.1109/JIOT.2021.3110412

[15] GHOLAMI A, BARAS J S. Collaborative cloud-edge-local computation offloading for multi-component applications [C]//2021 IEEE/ACM Symposium on Edge Computing (SEC). IEEE, 2021: 361-365. DOI: 10.1145/3453142.3493515

[16] SEDGHANI H, FILIPPINI F, ARDAGNA D. A randomized greedy method for AI applications component placement and resource selection in computing continua [C]//Proceedings of 2021 IEEE International Conference on Joint Cloud Computing. IEEE, 2021: 65-70. DOI: 10.1109/JCC53141.2021.00022

[17] CHEN X W, TANG S T, LU Z H, et al. iDiSC: a new approach to IoT-data-intensive service components deployment in edge-cloud-hybrid system [J]. IEEE access, 7: 59172-59184. DOI: 10.1109/ACCESS.2019.2915020

[18] FATICANTI F, DE PELLEGRINI F, SIRACUSA D, et al. Throughput-aware partitioning and placement of applications in fog computing [J]. IEEE transactions on network and service management, 2020, 17(4): 2436-2450. DOI: 10.1109/TNSM.2020.3023011

作 者 简 介

周吉喆，中国信息通信研究院工程师；主要研究方向为算网融合、通感算一体化；发表论文10余篇。

杨思远，中国信息通信研究院移动通信创新中心MTNet实验室主任工程师；长期从事移动通信领域的技术研究、技术试验等工作，负责多项LTE、NB-IOT基站等标准的制定工作，参与多项国家重大专项课题。

王志勤，中国信息通信研究院副院长、中国通信标准化协会无线通信技术工作委员会主席、中国通信学会无线及移动通信委员会主任委员，教授级高级工程师，“新一代宽带无线移动通信网”国家科技重大专项副总工程师；主要研究方向为无线移动通信技术和标准。