



5G 网络中 NSA 控制面和用户面时延性能分析

张丹¹, 王磊¹, 王晓琦¹, 席思雨²

(1. 中国移动通信集团陕西有限公司, 陕西 西安 710077;

2. 中国移动通信集团设计院有限公司, 北京 100080)

摘要: 不同于 2G/3G/4G 的独立组网, 5G 的网络架构分为两种: 非独立 (NSA) 组网和独立 (SA) 组网。NSA 作为非独立架构, 能实现 5G 的快速部署。低时延是 5G 非常重要的特点, 其实现需要一系列技术的有机结合。针对 2.6 GHz NR 帧结构, 结合理论分析和实际案例数据, 研究了 NSA 架构下控制面时延和用户面时延的现状和问题: 当前, 商用芯片级别的终端控制面时延约为 359~510 ms; 开启预调度功能时, NSA 好点空口双向用户面时延为 11~14 ms, 关闭预调度后, 用户面时延增加至 15~17 ms。为了进一步优化控制面时延性能, 可采用设置 NSA 的锚点与 NR 子帧同步、将 B1 的 *timetottrigger* 设置得略小等方案; 而采取 SR 周期自适应方式则可以降低用户面时延。并给出了时延优化方案, 给 NSA 网络部署提供了相应的思路和建议。

关键词: 5G; 帧结构; 控制面时延; 用户面时延

中图分类号: TN929.5

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2020260

Performance analysis of NSA control plane and user plane delay in 5G network

ZHANG Dan¹, WANG Lei¹, WANG Xiaoqi¹, XI Siyu²

1. China Mobile Group Shaanxi Co., Ltd., Xi'an 710077, China

2. China Mobile Group Design Institute Co., Ltd., Beijing 100080, China

Abstract: Different from 2G/3G/4G, the 5G network architecture contains two parts: NSA (non-standalone) and SA (standalone) networking. NSA can help network deploys faster. Low latency is a very important feature of 5G, and its implementation requires a combination of technologies. Based on the 2.6 GHz NR frame structure, combined with theoretical analysis and real case, the current situation and problems of control plane delay and user plane delay were discussed, and the delay optimization method was provided, which provided the corresponding ideas and suggestions for NSA network deployment.

Key words: 5G, frame structure, control plane delay, user plane delay

1 引言

不同于 2G/3G/4G 的独立组网, 5G 的网络架

构分为两种: 非独立 (NSA) 组网和独立 (SA) 组网。SA 作为独立架构, 能够一步到位实现 5G 的功能, 但是存在端到端成熟度风险^[1]。NSA 则



是 4G/5G 紧耦合，5G 依附于 4G 基站进行工作，无法独立组网，但能在初期实现 5G 快速部署。5G 时代降低时延可以降低基站和终端的响应时间，有效提高用户感知，时延是 5G 非常关键的性能。本文基于 NSA 组网架构和 2.6 GHz NR 帧结构，研究了控制面时延和用户面时延的基本原理和实际案例数据，指出了相应的问题并给出了时延优化方案，给 NSA 网络部署提供了相应的思路和建议。

2 2.6 GHz 帧结构方案分析

2.1 2.6 GHz 帧结构候选方案分析

在 5G eMBB 场景下，按照 30 kHz 的子载波间隔，有典型的几种帧结构配置：2.5 ms 双周期、2.5 ms 单周期、2 ms 单周期、5 ms 单周期等。

2.5 ms 双周期帧结构如图 1(a)所示，每 5 ms 包含 5 个全下行时隙、3 个全上行时隙和 2 个特殊时隙。其中，存在两个连续的上行时隙 (UL slot)，可以发送长 PRACH 格式，有利于提升上行覆盖能力，缺点是双周期实现比较复杂。

2.5 ms 单周期帧结构如图 1(b)所示，每 2.5 ms 包含 3 个全下行时隙、1 个全上行时隙和 1 个特殊时隙，单周期实现方式相比于双周期更简单，有更多下行时隙，有利于提升下行吞吐量，缺点是无法配置长 PRACH 格式。

2 ms 单周期帧结构如图 1(c)所示，包含 2 个全下行时隙、1 个全上行时隙和 1 个特殊时隙，该配置可以有效减少调度时延。

5 ms 单周期的配置如图 1(d)所示，每 5 ms

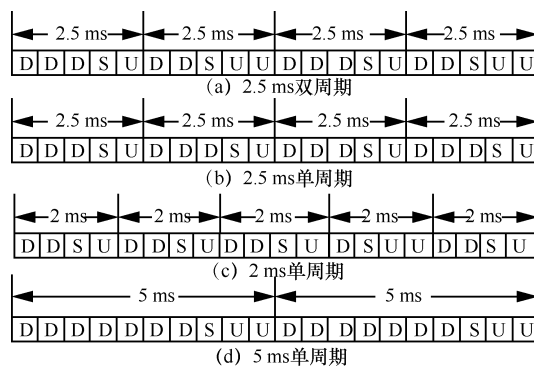


图1 典型帧结构配置

包含 7 个全下行时隙、2 个上行时隙和 1 个特殊时隙，2.6 GHz LTE 也采用了该配置。

2.2 2.6 GHz 帧结构配置方案建议

考虑 2.6 GHz NR 与 LTE 共存的原因，本文采用 5 ms 单周期作为 2.6 GHz 频段的帧结构方案。

根据 3GPP 规范，NR 的下行邻道泄露比为 45 dB，在 2.6 GHz NR 与 2.6 GHz TDD 邻频共存的情况下，如果上下行时隙不能对齐将出现严重的上下行反向干扰，所以 NR 的上下行时隙必须与 LTE 2.6 GHz 对齐。基于此原因，可选择 5 ms 单周期 Pattern-1 或 Pattern-2 作为 2.6 GHz NR 的帧结构，如图 2 所示。两种方案优势是实现简单，缺点是帧头需有一定的偏移，Pattern-1 帧头需推迟 3 ms，Pattern-2 帧头需提前 2 ms。

3 NR 时延理论分析

3.1 NSA 控制面时延理论分析

NSA 控制面时延信令流程如图 3 所示，为方便分析，将时延信令流程分为以下 4 段。

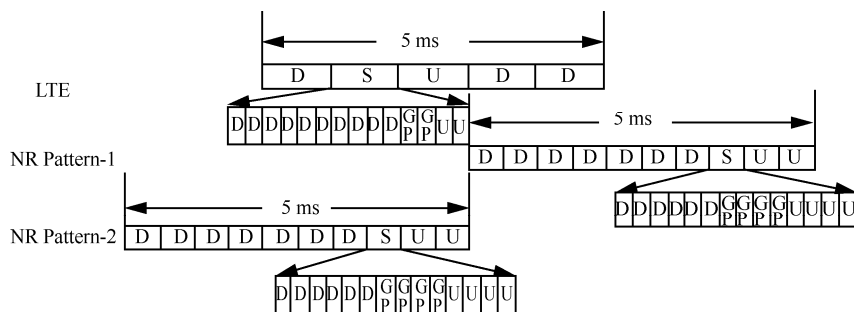


图2 5 ms 单周期帧结构

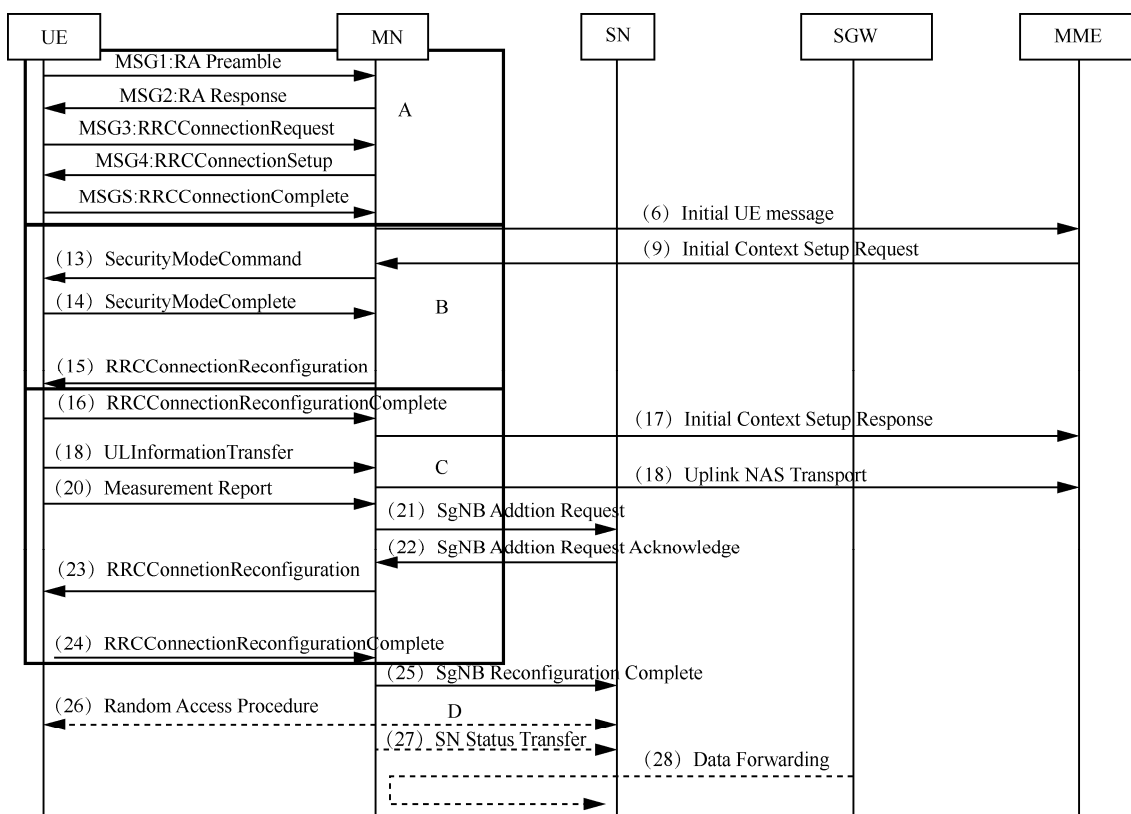


图 3 NSA 控制面时延信令流程

(1) 控制面时延 A 是 LTE 随机接入的时延, 从终端在 LTE 发出 Preamble(MSG1)到发出 RRC 连接成功 (MSG5)。时延 A 取决于 LTE, 一般为 40 ms 级别。

(2) 控制面时延 B 是 LTE 建立承载的时延, 从终端在 LTE 发出 Preamble(MSG1), 到收到“第一条 RRC 重配 (含 5G 测控)”消息。在 4G 时代, 时延 B 取决于 LTE, 一般为 100 ms 级别, 与 4G 不同的是, NSA 的时延 B 增加了 NR 终端能力的查询与上报, 其中包含 NR 单独的能力和 EN-DC 双连接能力。

(3) 控制面时延 C 是 5G 测量加配置双连接的时延, 从终端收到“第一条 RRC 重配 (含 5G 测控)”消息, 到终端在 LTE 发出“第二条 RRC 重配 (包含 5G Scell 添加) 完成”之间的时延。时延 C 与帧结构的选择和终端的测量实现都有关系, 这段包含测量配置、异频 GAP 测量、SCG 添

加及重配置过程。

(4) 控制面时延 D 是终端在 5G Scell 接入的时延, 非竞争模式下, 从终端在 5G 发出 Preamble (MSG1) 到收到 MSG2 为止。

3.2 SA 控制面时延理论分析

相比 NSA, 整体时延不包括 4G 随机接入到承载建立、5G 测量与辅载波添加等流程。

Idle 态时延为 SA 测试终端从 Idle 态到 DRB 建立完成, 包含 5G 随机接入 MSG1-MSG5 和安全激活、RRC 重配过程 (UE 发出第一条重配完成消息时结束)。

Inactive 态时延为 SA 终端从 Inactive 态到 DRB 恢复完成, 包含 5G 随机接入 MSG1-MSG5 (resume 完成) 的过程。

3.3 NR 用户面时延理论分析

NR 上下行传输流程见表 1^[2-3]。下行传输分为初传和重传两部分, 步骤 1~步骤 4 为初传流程,



步骤 5~步骤 13 为重传流程。

本文中，上行传输分为关闭预调度和开启预调度两个部分，关闭预调度也叫动态调度。当 UE 需要发送数据时，会通过 PUCCH 信道发送 SR 请求，基站收到请求后，会给 UE 发送 UL Grant Count 分配上行资源；开启预调度时，调度器始终为 UE 上行数据分配固定的资源，不需要 UE 发起调度请求就可以直接发送数据分组。

- 关闭预调度的上行传输分为调度请求、初传和重传 3 个部分，步骤 1~9 为调度请求，步骤 10~13 为初传流程，步骤 14~22 为重传流程。
- 开启预调度分为初传和重传两个部分，步骤 1~4 为初传流程，步骤 5~12 为重传流程，开启预调度的上行传输与关闭预调度的上行传输的区别在于少了调度请求这个部分，它可

以有效地减少用户面时延，但是当 UE 需要传输的资源较少时，会引起额外的资源消耗。

用户面时延和帧结构关系较大，用户面时延主要包含高层时延、物理层初传时延、物理层重传时延、传输时延和核心网时延等。

本文以 5 ms 单周期帧结构作为分析基础，假设数据在某 symbol 到达，按照表 1 的步骤完成初传和重传，考虑数据可能在任意 symbol 到达，可以将帧结构内每个 symbol 作为起始点，得到所有时延的列表，然后统计该列表的平均值。在表 2 的统计结果中，设置重传概率为 10%，总体时延由高层时延、90%的初传时延和 10%的重传时延组成。

NSA 用户面时延基本流程见表 2。在开启免调度的场景下，UE 到基站的环回时延理论平均值为 6 ms（此计算仅包含无线时延，不包含传输和核心网时延），在关闭预调度的场景下，UE 到基

表 1 NSA 上下行传输流程

步骤序号	步骤说明		步骤说明		步骤说明	
	下行		上行-关闭预调度		上行-开预调度	
1	基站编码、调制（用户数据）	初传	UE 编码、调制（SR）	调度	UE 编码、调制（用户数据）	初传
2	等待下行传输机会		等待上行传输机会		等待上行传输机会	
3	下行传输（用户数据）		上行传输（SR）		上行传输（用户数据）	
4	UE 解码（用户数据）		基站解调（SR）		基站解调（用户数据）	
5	UE 编码、调制（NACK）	重传	编码调制（UL Grant）		编码调制（UL Grant）	重传
6	等待上行传输机会		MAC 调度		MAC 调度	
7	上行传输（NACK）		等待下行传输机会		下行传输（UL Grant）	
8	基站解密（NACK）		下行传输（UL Grant）		UE 解码（UL Grant）	
9	MAC 调度		UE 解码（UL Grant）		UE 编码、调制（用户数据）	
10	基站编码、调制（用户数据）		UE 编码、调制（用户数据）	初传	等待上行传输机会	
11	等待下行传输机会		等待上行传输机会		上行传输（用户数据）	
12	下行传输（用户数据）		上行传输（用户数据）		基站解密（用户数据）	
13	UE 解码（用户数据）		基站解密（用户数据）			
14			编码调制（UL Grant）	重传		
15			MAC 调度			
16			等待下行传输机会			
17			下行传输（UL Grant）			
18			UE 解码（UL Grant）			
19			UE 编码、调制（用户数据）			
20			等待上行传输机会			
21			上行传输（用户数据）			
22			基站解密（用户数据）			

表 2 NSA 用户面时延基本流程

类别	下行	上行-关预调度	上行-开预调度
gNB 高层 (symbol)	7	7	7
UE 高层 (symbol)	7	7	7
物理平均时延 (仅初传, symbol)	35.7	228.5	81.9
物理平均时延 (含重传, symbol)	136.5	368.5	220.5
平均总时延 (symbol)	59.8	256.5	109.8
平均总时延 (ms)	2.1	9.2	3.9

站的环回时延理论平均值为 11.3 ms。如果核心网在本地,基站到核心网之间的传输时延为 1~3 ms;如果核心网在邻省,传输时延为 4~8 ms。在实际场景中,终端还需要 2~10 ms 的处理时延,gNode B 设备处理能力不同相应的处理时延也不一样。

4 NSA 控制面时延分析与建议

4.1 NSA 控制面时延实际案例数据分析

选择室外密集城区主测 NSA 的小区,周围站点开启且空扰,CU-DU 采用合设,关闭预调度的场景,用一部 NSA 测试终端进行控制面时延的测试。测试方法为将 NSA 测试终端放置在小区好点,发起业务(如 ping 分组)触发从 idle 态到 NR 连接态,重复 50 次,记录控制面时延 A、B、C、D 段。

NSA 控制面时延在实际测试中的 5 类案例如图 4 所示,案例 1 采用 TUE 进行测试,案例 2~案例 5 采用两款不同类型的商用芯片进行测试。根据图 4 中的测试结果进行统计,TUE 级别在从 LTE 发起随机接入,到 NR 完成随机接入的控制面时延为 276 ms 左右,商用芯片级别的终端控制面时延约为 359~510 ms。从结果可以看出:

(1) A 段时延平均为 24~44 ms, LTE RRC 建立平均时延为 50 ms 以内,时延 A 测试结果符合预期;

(2) 时延 B 芯片级测试结果约为 82~142 ms,

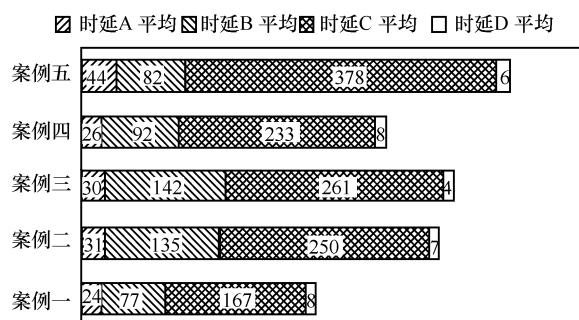


图 4 NSA 控制面时延案例

TUE 级别为 77 ms,可以看出不同的案例时延 B 的差别较大,造成此类结果的主要原因是时延短的案例只有 1 次 UE 能力查询,时延长的案例一般有 2~4 次能力查询,可通过减少 UE 能力查询次数进一步优化时延;

(3)时延 C 芯片级测试平均时延为 233~378 ms, TUE 级别约为 166 ms,相比较于成熟商用终端, TUE 级别可通过采取提前启动 5G 同步,缩短 4G/5G 交互时间等措施缩短控制面时延,此优化方案在 TUE 中实现简单。

(4) 时延 D 芯片级测试平均时延为 4~8 ms,符合预期。

4.2 NSA 控制面时延优化建议

NSA 控制面时延相比 LTE 控制面时延增加约 200~400 ms,用户接入体验略有降低,优化控制面时延主要有以下 5 种方案^[4]。

(1)网络侧可通过设置 NSA 的锚点与 NR 子帧同步来降低 C 段时延。终端有 5 ms 的测量窗口,如



果子帧对齐，就能确定 SSB 所在的子帧位置，减少搜索时间。该方案的优势是简单易实现，缺点是部分厂商在 1.8/1.9 GHz 时期无此需求，因此无法通过锚点与 NR 子帧同步来完成此项功能。经外场测试，子帧同步后，C 段时延能降低 150 ms 左右。

(2) 对于 NSA 的初始接入，为了快速添加 SN，可以将 B1 的 `timetotrieger` 设置得略小一些，该参数指当前服务小区涉及重选的时间滞后量，当参数设置的过大，会导致小区重选在应该执行的时候没有机会执行，当参数设置的过小，可能会导致频繁发起小区重选。未来可将 `timetotrieger` 自适应设置方式作为优化方向，当 NR 信号稳定可通过缩短 `timetotrieger` 来优化时延，当 NR 信号弱时可通过延长 `timetotrieger` 避免乒乓切换。

(3) 网络侧可通过异频测量时不配置 GAP 的方式进一步缩短 C 段控制面时延。无 GAP 测量既可以在测量期间传送数据，同时能有效减少控制面 C 段时延 100 ms 左右，目前大部分终端均支持不配置 GAP 的 B1 测量。

(4) 通过优化 UE 能力查询次数降低控制面 B 段时延，从目前的调研结果发现，不同的案例控制面 B 段时延结果不一样的主要原因是 UE 能力查询次数不一样。经外场测试，如果降低查询次数，B 段时延能降低 1.4~2.2 倍。

(5) 通过打开预调度开关降低 `msg1~msg5` 的时延，在关闭预调度的场景中，UE 上行数据传输需要先等待调度请求 (SR) 的发送机会。在发送 SR 后，网络侧给 UE 发送上行调度授权，然后 UE 才能发送数据。NR 默认 SR 周期为 40 slot，每 20 ms 才有一次发送机会。这将大幅度增加上行时延。在预调度场景免除了 SR 等待以及 SR 调度的时延，所以可以大幅降低控制面时延。经外场测试，打开预调度开关后，MSG1~5 的时延均值由 35~40 ms 降低到 20 ms 左右。

相较于 4G 和 SA，NSA 的控制面时延有其独特的特点：5G 帧结构设计带来的调度时序和

HARQ 时序相较于 4G 有一定优势；整体时延包括 4G 随机接入到承载建立、5G 测量与辅载波添加等流程，相较于 SA 控制面时延整体延长了 260~400 ms。

5 NSA 用户面时延分析与建议

5.1 NSA 用户面时延实际案例数据分析

选择室外密集城区主测 NSA 的小区，周围站点开启且空扰，CU-DU 采用合设，网络关闭分流的场景，用一部 NSA 测试终端进行用户面时延的测试。测试方法为：

步骤 1 将 NSA 测试终端放置在小区好点，向核心网服务器发起 2 000 Byte 大小的 Ping 分组，重复 100 次；

步骤 2 基站向核心网服务器发起 2 000 Byte 大小的 Ping 分组，或者核心网向基站服务器发起 2 000 Byte 大小的 Ping 分组，重复 100 次，并记录 100 次 Ping 分组时延数据。

NSA 用户面时延在实际测试中的 4 类案例如图 5 所示。从图 5 可以看出，在现有 5 ms 帧结构下，开启预调度功能时，案例 1~案例 4 的 NSA 好点空口双向时延为 11~14 ms，关闭预调度后，案例 1~案例 3 的时延为 15~17 ms，案例 4 的时延为 24.8 ms。从测试结果可以看出：

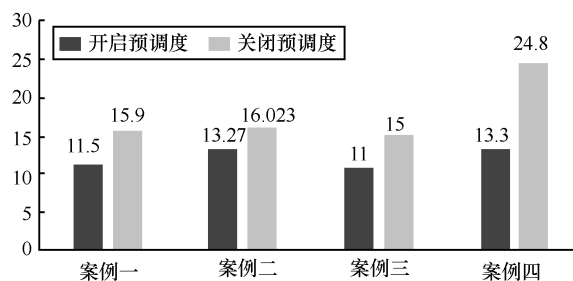


图 5 NSA 用户面时延案例

(1) 核心网的位置是影响用户面时延的关键因素，在测试过程中发现，本地核心网时延一般为 1 ms 左右，跨城市核心网约为 6 ms，跨省份约为 26 ms。

(2) 空口时延长短与分包时延有较大的关系, SR 上报不及时、SR 调度偏小会导致分片、重传较多, 分组后约增加 5 ms 空口时延。案例 4 关闭预调度后的时延超过了 20 ms 主要是进行了分组传输。

(3) 根据理论预期, 开启预调度相较于关闭预调度可以缩短空口时延 5 ms 左右, 从图 5 可以看出, 开启预调度确实可以达到缩短时延的效果。

5.2 NSA 用户面时延优化建议

在未开启预调度的场景下, NSA 的用户面时延相较于 LTE 平均缩短 9~16 ms, 理论上在相同场景下, NR 的用户面时延低于 LTE 的用户面时延, 得益于 NR 调度等待和处理时延更短等因素。NSA 的用户面有以下措施可以进一步降低用户面时延^[5]。

(1) 基站可以通过优化调度资源使得在传输数据分组时不产生分包, 经外场测试, 不分包时, 空口双向时延可降低约 5 ms。

(2) 采取 SR 周期自适应的方式, 当用户数较少时, 可以使用短 SR 周期, 降低时延; 当用户数较多时, 可以使用长 SR 周期, 提升容量。

6 结束语

本文基于 NSA 组网架构, 首先给出了 2.6 GHz 帧结构配置建议, 然后研究了控制面时延和用户面时延的基本原理, 最后在不同的场景下开展了大量的测试, 通过实际案例数据分析了 NSA 控制面时延和用户面时延的性能, 并给出了时延优化建议, 给 NSA 网络部署提供了相应的思路和建议。后续将进一步研究 SA 的控制面时延和用户面时延, 并和 NSA 的时延性能进行对比。

参考文献:

- [1] 刘玮, 董江波, 任冶冰. 面向商用的 5G 网络关键问题研究及验证[J]. 电信科学, 2018, 34(8): 56-61.
LIU W, DONG J B, REN Y B. Research and verification of key issues in 5G network[J]. Telecommunications Science, 2018, 34(8): 56-61.
- [2] 刘玮. 5G NSA 分流与移动性方案分析与优化[J]. 移动通信,

2019(7).

LIU W. Analysis and optimization of 5G NSA shunting and mobility scheme[J]. Mobile Communications, 2019(7).

- [3] 方绍湖, 李馨, 卜斌龙. 基于开放平台小基站的 5G 数字室分解决方案[J]. 电信科学, 2019, 35(7): 69-77.
FANG S H, LI X, BU B L. 5G digital wireless indoor coverage solution based on open platform small cell[J]. Telecommunications Science, 2019, 35(7): 69-77.
- [4] 柏青, 吕平宝, 王勇, 等. 下一代网络时延提升的关键技术[J]. 电信快报, 2017(1): 31-34.
BAI Q, LV B P, WANG Y, et al. The key technology of next generation network delay enhancement[J]. Telecommunications Information, 2017(1): 31-34.
- [5] 张铁, 侯雪颖, 夏亮, 等. 5G 系统用户面时延浅析[C]//5G 网络创新研讨会(2018)论文集, 2018.
ZHANG Y, HOU X Y, XIA L, et al. Brief analysis of 5G system user face delay[C]//Workshop on 5G Network Innovation (2018), 2018.

[作者简介]



张丹(1973-), 女, 中国移动通信集团陕西有限公司工程师, 主要从事 4G/5G 方面的网络规划、优化工作。



王磊(1987-), 男, 中国移动通信集团陕西有限公司工程师, 主要从事 4G/5G 方面的网络规划、优化工作。



王晓琦(1982-), 男, 中国移动通信集团陕西有限公司工程师, 主要从事移动通信网络的规划、优化工作。

席思雨(1992-), 女, 现就职于中国移动通信集团设计院有限公司, 主要研究方向为移动通信新技术。