中国科学 : 信息科学 2025 年 第 55 卷 第 10 期: 2491–2500
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
面向低空经济的低空网络技术创新与应用专题 · 论文

# 面向具备智能在轨服务功能的 NTN-IoT 的上行帧资源部署方案设计

王雨[1], 李志强[1], 韩帅[1*], Abderrahim BENSLIMANE[2], 李成[3]

1. 哈尔滨工业大学电子与信息工程学院, 哈尔滨 150001, 中国
2. Laboratoire Informatique d'Avignon, University of Avignon, Avignon 84911, France
3. School of Engineering Science, Simon Fraser University, Burnaby V5A, Canada
* 通信作者. E-mail: hanshuai@hit.edu.cn

收稿日期: 2025–02–14; 修回日期: 2025–07–07; 接受日期: 2025–09–16; 网络出版日期: 2025–10–13

国家自然科学基金 (批准号: 62371166) 资助项目

**摘要** 本文面向具备智能在轨服务功能的 NTN-IoT 网络, 设计了一种基于帧结构资源部署优化方案. 多颗具备在轨推理功能的卫星服务众多地面用户, 这些地面用户主要产生常规通信数据和少量 AI 推理数据, 因此卫星需要合理安排这些数据的帧结构. 本文按照 NPUSCH (窄带物理层上行链路共享信道) 给出的帧结构特征对这一问题进行表述, 并进一步将这个问题描述为 0-1 背包问题. 为了资源部署效率, 本文设计了一种基于分支定界的方法, 该方法先对用户的通信信息进行处理然后再分配 AI 数据. 仿真结果表明, 本文的方案能够提升频谱效率, 这为海量用户的接入提供了优势, 同时本方案还能够对 AI 数据进行合理放置以降低平均推理时间.

**关键词** NPUSCH, IoT, 帧结构, 海量接入, 在轨推理

## 1 引言

随着新一代通信技术演进, 大规模低轨卫星网络已成为通信领域的研究热点. 低轨卫星凭借轨道高度低的特性, 其覆盖范围显著超越传统地面网络; 然而, 卫星运动的高动态性、长距离传输带来的时延挑战, 以及有限载荷资源的约束, 限制了卫星网络的大规模实际应用. 3GPP 组织在 38.821 标准中 [1] 明确了非地面网络 (NTN) 技术的两大发展方向: 面向地面物联网场景的 NTN-IoT 以及针对空间通信场景的 NR-NTN. 值得注意的是, 38.821 规范中定义的 NTN-IoT 初期以地球同步轨道 (GEO) 卫星为服务主体, 而近年来兴起的手机直连卫星技术为 NTN-IoT 开辟了全新技术路径 —— 该技术有望使海量地面用户突破地面基础设施限制, 实现基于低轨卫星 (LEO) 的大规模物联网接入. 基于上述背景, 本文聚焦具备在轨智能服务能力的 NTN-IoT 网络, 研究在海量用户上行传输场景下, 如何优化帧结构资源部署效率, 同时解决常规通信数据与并发 AI 推理数据的协同适配问题.

**引用格式:** 王雨, 李志强, 韩帅, 等. 面向具备智能在轨服务功能的 NTN-IoT 的上行帧资源部署方案设计. 中国科学: 信息科学, 2025, 55: 2491–2500, doi: 10.1360/SSI-2025-0070
Wang Y, Li Z Q, Han S, et al. Uplink frame resource placement scheme design for NTN-IoT with intelligent on-orbit service function. Sci Sin Inform, 2025, 55: 2491–2500, doi: 10.1360/SSI-2025-0070

 

如文献 [2] 设计了星基物联网 (S-IoT) 系统, 该系统通过集成 AIS 船舶监测、VDES 双向通信、ADS-B 航空监视、DCS 传感器数据采集及 ESR 紧急搜救的 5S 载荷 (符合 1.5U 立方星标准), 实现广域覆盖与多功能服务, 其设计在天拓五号卫星 (2020 年发射) 经 3 年在轨验证, 支持海洋、航空及应急场景应用, 兼顾紧凑性与高效部署. 直接卫星物联网 (DtS-IoT) 利用低轨卫星作为网关, 解决偏远地区物联网数据传输难题, 文献 [3] 提出基于 LoRa 和 LR-FHSS 的方案, 结合卫星轨迹优化上行传输策略, 仿真表明其可倍增节点数量并提升扩展性, LR-FHSS 扩展性提升 75 倍但功耗增加 30%. 文献 [4] 研究利用低轨卫星为地面用户设备提供窄带物联网 (NB-IoT) 连接, 针对传统资源分配算法不适用于卫星系统的局限性, 提出一种新上行资源分配策略, 综合考虑信道变化、卫星覆盖时间和用户需求, 为未来研究提供框架. 文献 [4] 面向 NTN-IoT 的上行帧结构, 给出了一种基于分支定界的帧结构优化方案, 这一方案为我们这篇文章带来了很大启发.

近年来, 卫星在轨应用 AI 技术也逐渐成为研究者关注的热点之一. DeepSeek-R1 [5] 基于 DeepSeek-V3 的 MoE 架构, 包含 16 个专家网络, 每个专家专注于特定领域 (如数学、代码、逻辑等), 通过动态路由机制激活相关专家. 模型总参数量为 671 B, 但实际激活参数量仅 37 B, 显著降低计算成本. 这一部署方式为在空间网络应用 AI 技术提供了可能.

文献 [6] 针对非静止轨道卫星 (NGSO) 与静止轨道卫星 (GSO) 共享频段引发的干扰问题, 提出基于生成式 AI (VAE 和 Transformer 检测器 TrID) 的时域信号检测方法, 实验表明 TrID 检测精度较传统方法提升 31.23%, 为频谱管理提供创新方案. 文献 [7] 提出星基全球遥感框架, 整合在轨云计算与 AI, 利用大规模低轨星座实现实时遥感及场景分析, 无需回传数据, 支持地球资源监测、灾害救援、智慧城市等应用, 兼具高效、绿色与低成本特性. 面向 6G 卫星通信网络大规模发展的建模与干扰难题, 文献 [8] 提出了生成式 AI 代理建模及 MoE-PPO 策略, 结合 LLM 交互与专家知识设计传输方案, 实验验证模型准确性且 MoE-PPO 优于基准方法并具备广泛适应性. 文献 [9] 研究了低轨卫星的物理层认证, 发现 CNN 和自编码器可有效认证卫星换能器, 但受卫星链路低带宽限制, 可能无法检测所有欺骗攻击.

## 2　系统模型

AI 在轨应用基础已经具备充分的应用条件, 因此本文对多种数据协同方式调度难题, 考虑多颗卫星服务协同服务多种地面用户的场合, 系统模型如图 1 所示. 为了方便说明, 我们给出本文涉及到的所有参数及其说明, 如表 1 所示.

在我们设定的场景中, 不同的卫星能够处理不同的 AI 信息, 但是一颗卫星只能处理一种特定类型的 AI. 所有覆盖范围内的卫星均能够接收常规通信数据. AI 数据的业务类型可以使用一组连续的整数进行表示, 如下所示:

$$A = (1, 2, \ldots, k),\ k \in \mathbb{Z}. \tag{1}$$

用户和卫星均有对应的数据标识类型, 我们假设卫星能够处理与它自身数据标识不同的业务, 但是数据标识之间的差距影响 AI 数据的处理效果, 差距越小, 意味着业务类型越接近, 数据处理的效率越高. 因此我们可以使用一种加权的数据量来表示 AI 数据的接受效果.

$$\delta_{ij} = \|s_i - u_j\|,\ s_i \in A,\ u_j \in A, \tag{2}$$

其中 $s_i$ 表示第 $i$ 颗卫星的数据标识类型, $u_j$ 表示第 $j$ 个用户的数据标识类型. 在 NPUSCH 中, 规定了用户上传的数据的大小和放置的规则, 这些数据块占据特定大小的带宽和时隙. 对于海量用户介入的场景而言, 用户相对卫星的距离和相对移动速度都不等, 因此 3GPP 设定了根据用户通信状态自适应调制的机制.

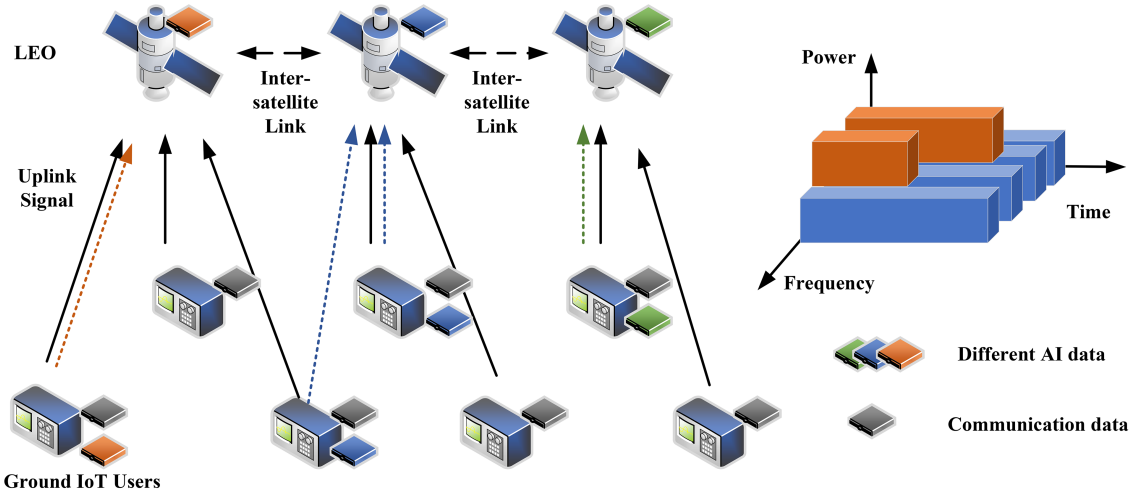


图 1　(网络版彩图) 系统模型.

**Figure 1**　(Color online) System model.

表 1　系统模型的参数.

**Table 1**　Parameter for system model.

| Parameter | Representation | Meaning |
|---|---|---|
| User index | $i$ | Represents the $i$-th user, total of $I$ users |
| Satellite index | $j$ | Represents the $j$-th satellite, total of $J$ satellites |
| Total system bandwidth | $B$ | Represented by the number of subcarriers |
| Total number of time slots in the system | $W$ | Represented by the number of time slots |
| Time-domain resource index | $p, r$ | Belongs to the set $\mathcal{X} = 1, 2, \ldots, M$ |
| Frequency-domain resource index | $q, s$ | Belongs to the set $\mathcal{Y} = 1, 2, \ldots, N$ |
| Conventional data block time-domain length | $w_i$ | Number of time-domain resource units occupied by user $i$'s conventional data block |
| Conventional data block frequency-domain width | $b_i$ | Number of frequency-domain resource units occupied by user $i$'s conventional data block |
| AI data block time-domain length | $w_i^{ai}$ | Number of time-domain resource units occupied by user $i$'s AI data block |
| AI data block frequency-domain width | $b_i^{ai}$ | Number of frequency-domain resource units occupied by user $i$'s AI data block |
| Conventional data block value | $P_i$ | Value of user $i$'s conventional data block |
| AI data block value | $P_i^{ai}$ | Value of user $i$'s AI data block |
| Conventional data block placement decision | $z_{ipq}$ | Indicates whether user $i$'s conventional data block is placed on resource grid $(p, q)$ |
| AI data block placement decision | $z_{ip'q'}$ | Indicates whether user $i$'s AI data block is placed on resource grid $(p', q')$ |
| Satellite service type | $s_i$ | Identifier of the AI service type that satellite $i$ can process |
| User service type | $u_j$ | Identifier of the AI data service type sent by user $j$ |
| Service type matching weight | $\delta_{ij}$ | Represents the service type difference between user $i$ and satellite $j$, usually $\lvert s_j - u_i \rvert$ |

本文假设低轨卫星的运行轨道是固定的, 地面 IoT 用户也是固定的, 信道能够完美均衡, 用户主要根据自身的链路衰减来确定自适应调制的等级. NPUSCH 的资源块一般布置形式可以描述为图 2. 决定资源块放置的因素主要有资源块的大小即时频资源占用量和位于帧当中的位置. 具体地,

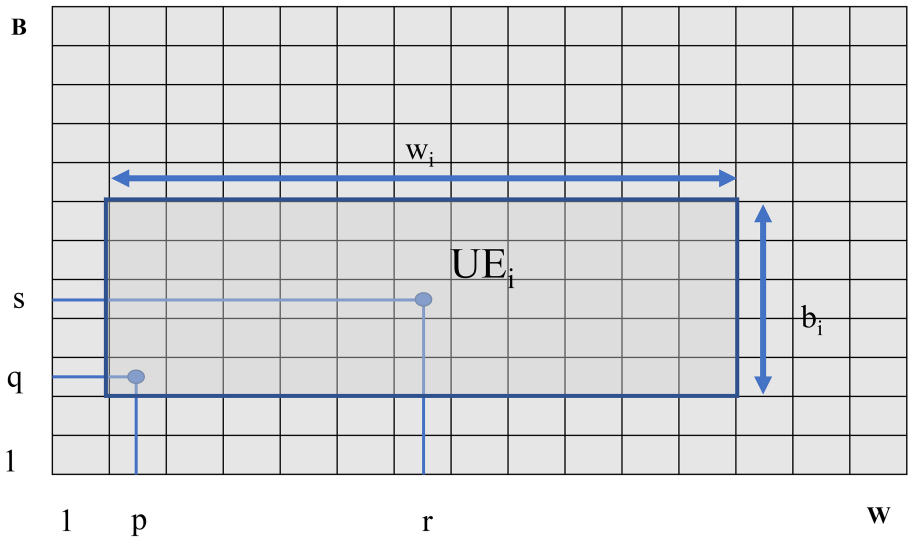


**图 2　(网络版彩图) NPUSCH 的资源块一般布置形式.**

**Figure 2**　(Color online) Resource placement of NPUSCH.

用户 $i$ 的常规数据块可表征为一个矩形, 其尺寸由所需时域资源单元数 $w_i$ 和频域资源单元数 $b_i$ 决定, 该矩形放置于帧中的位置由其左下角坐标 $(p,q)$ 确定, 其中 $p\in\mathcal{X}=\{1,2,\ldots,M\}$ 为时域索引, $q\in\mathcal{Y}=\{1,2,\ldots,N\}$ 为频域索引. 此外, 为了描述资源块承载数据的传输效率, 我们使用 $P_i$ 表示用户 $i$ 的常规数据块的价值, 该价值由用户的自适应调制阶数决定. 对于用户产生的 AI 数据, 我们类似地定义了其数据块的价值 $P_i^{ai}$. AI 数据块同样需占用一定的时频资源, 其尺寸由参数时域长度 $w_i^{ai}$ 和频域宽度 $b_i^{ai}$ 决定, 该块在帧中的位置由其左下角坐标 $(p',q')$ 确定. 用户 $i$ 的常规数据与 AI 数据的资源分配决策分别由二元变量 $z_{ipq}$ 和 $z_{ip'q'}$ 表示, 若资源块被放置于对应位置则取值为 1, 否则为 0.

通过将所有用户的上传的数据块价值加和所得到的总数据价值进行最大化处理, 可以使我们的总目标是最大化加权多卫星共同收集的上行有效数据量. 可以将优化问题表示为式 (3) 的形式:

$$\max_{Z}\sum_{i=1}^{N}\sum_{p\in X}\sum_{q\in Y}P_i\cdot z_{ipq}+\sum_{i=1}^{N}\sum_{p'\in X}\sum_{q'\in Y}P_i^{ai}z_{ip'q'},\tag{3}$$

其中

$$\sum_{i=1}^{N}\sum_{\{p\in X|r-w_i+1\leqslant p\leqslant r\}}\sum_{q\in Y}b_i\cdot z_{ipq}\leqslant B,\quad\forall r\in X,\tag{4}$$

$$\sum_{i=1}^{N}\sum_{\{p'\in X|r-w_i+1\leqslant p'\leqslant r\}}\sum_{q'\in Y}b_i\cdot z_{ip'q'}\leqslant B,\quad\forall r\in X,\tag{5}$$

$$\sum_{i=1}^{N}\sum_{p\in X}\sum_{\{q\in Y|s-b_i+1\leqslant s\}}w_i\cdot z_{ipq}\leqslant W,\quad\forall s\in Y,\tag{6}$$

$$\sum_{i=1}^{N}\sum_{p'\in X}\sum_{\{q'\in Y|s-b_i+1\leqslant q'\leqslant s\}}w_i\cdot z_{ip'q'}\leqslant W,\quad\forall s\in Y,\tag{7}$$

$$\sum_{i=1}^{N}\sum_{\{p\in X|r-w_i+1\leqslant p\leqslant r\}}\sum_{\{q\in Y|s-b_i+1\leqslant s\}}z_{ipq}\leqslant 1,\quad\forall r\in X,\quad\forall s\in Y,\tag{8}$$

$$\sum_{i=1}^{N}\sum_{\{p'\in X|r-w_i+1\leqslant p'\leqslant r\}}\sum_{\{q'\in Y|s-b_i+1\leqslant q'\leqslant s\}} z_{ip'q'}\leqslant 1,\quad \forall r\in X,\quad \forall s\in Y, \tag{9}$$

$$\sum_{p\in X}\sum_{q\in Y} z_{ipq}\leqslant 1, \text{for } i=1,2,\ldots,N, \tag{10}$$

$$\sum_{p'\in X}\sum_{q'\in Y} z_{ip'q'}\leqslant 1, \text{for } i=1,2,\ldots,N. \tag{11}$$

通过上述约束条件, 可以对资源分配问题的形式化表述有更清晰的理解. 其中, 约束与分别限制了在任意时域位置 $r\in\mathcal{X}$ 上, 所有用户常规数据与 AI 数据所占用的频域资源总宽度 (即频域单元数) 不得超过系统总带宽 $B$ (以子载波个数表示). 类似地, 约束与限制了在任意频域位置 $s\in\mathcal{Y}$ 上, 两类数据所占用的时域资源总长度 (即时域单元数) 不得超过系统一帧的总时隙数 $W$. 这 4 项约束共同确保了无论常规数据还是 AI 数据的放置, 都不会超出当前帧的时频资源总容量. 由于 AI 数据以非正交方式叠加在常规数据之上, 因此对其的资源约束是独立于常规数据单独施加的. 此外, 约束与保证了帧结构中的每个时频资源格至多只能被一个用户的常规数据或 AI 数据块占用, 从而避免了资源分配冲突. 约束与则规定每个用户最多只能接入一颗卫星为其传输常规数据, 以及另一颗卫星处理其 AI 数据, 即一颗卫星处理常规通信业务, 另一颗卫星处理 AI 推理业务, 体现了用户数据可被分拆至不同卫星服务的模型特性.

## 3　问题分析

我们将系统模型中提到的问题分为两个部分, 第一个部分是基础通信数据每帧价值最大化, 另一部分则是 AI 数据选择合适的卫星进行放置以提高处理效率. 为了增大资源的利用效率, 我们引入了非正交多址接入技术, 将通信数据和 AI 数据在功率域进行叠加, 以通信数据作为基础, AI 数据叠加在其上, 根据业务类型对 AI 数据进行合理的数据流向控制. 下面分别对这两个问题进行分析.

### 3.1　基础通信数据价值最大化问题

问题 (3) 是一个二维多要素背包问题, 这个问题解决难度极大, 因此需要对这个问题进行进一步的降维, 令数据块在每一个子载波上被分配. 因此为了使基础通信数据的价值最大化, 这部分问题可以重新进行表示:

$$\max_{z}\sum_{i=1}^{B}\sum_{k=1}^{N_j} P_k\cdot z_{ik} \tag{12}$$

$$\text{s.t. }\sum_{k=1}^{N} w_k\cdot z_{ik}\leqslant W_j,\ \text{for } i=1,2,\ldots, \tag{13}$$

$$\sum_{i=1}^{B} z_{ik}\leqslant 1,\ \text{for } k=1,2,\ldots,N_j, \tag{14}$$

$$z_{ik}\in\{0,1\},\ \text{for } i=1,2,\ldots,B,\ k=1,2,\ldots,N_j, \tag{15}$$

其中

$$z_{ik}=\begin{cases}1, \text{ 如果第 } k \text{ 个 UE 被分配到第 } i \text{ 个子载波},\\ 0, \text{ otherwise.}\end{cases} \tag{16}$$

### 3.2　AI 数据放置问题

在本文的场景假设中, 地面用户所产生的 AI 数据量要远小于通信数据的量级, 这些数据不会充满整个帧, 因此数据的尺寸限制不再是我们考虑的主要问题. 我们对 AI 数据进行处理的目的是最小化所有业务数据和目的卫星之间的业务类型差:

$$\min \sum_{i=1}^{N} \delta_{ij} z_{ip'q'} \tag{17}$$

$$\text{s.t. } A = (1, 2, \ldots, k),\ k \in \mathbb{Z}, \tag{18}$$

$$\sum_{i=1}^{N} \sum_{\{p' \in X | r - w_i + 1 \leqslant p' \leqslant r\}} \sum_{\{q' \in Y | s - b_i + 1 \leqslant q' \leqslant s\}} z_{ip'q'} \leqslant 1,\ \forall r \in X,\ \forall s \in Y, \tag{19}$$

以上约束分别用来表示业务类型的整数约束和放置标记的唯一性约束.

## 4　方案设计

在经过问题拆分后, 我们已经将原问题变为了一个二维单要素背包问题. 对于形如式 (12) 的二维背包问题而言, 当前的解决方案复杂度都很高, 只能适应用户数量很少的情况, 因此在本文所描述的海量用户接入的情况下必须要找到一种简便的方法才能有效地解决问题. 因此我们引入分支定界方案来处理, 这种方案可以将二维的背包问题降维, 从而降低算法的复杂程度. 对于 0-1 MKP 的精确解, 我们参考文献 [2] 中提出的方法, 其中提出了一种专门为求解大型问题而设计的精确算法实例.

(1) 推导上界: 该算法利用代理松弛来推导上界. 这样, 0-1 MKP 问题将转化为普通的 0-1 KP, 易于通过动态线性规划求解

$$\max_{z'} \sum_{j=1}^{N_j} P_k \cdot z'_k \tag{20}$$

$$\text{s.t. } \sum_{k=1}^{N_j} w_k \cdot z'_k \leqslant B \cdot W_j, \tag{21}$$

$$z'_k \in \{0,\ 1\},\ \text{for } k = 1, 2, \ldots, N_j, \tag{22}$$

其中引入的决策变量 $z'_k = \sum_{i=1}^{B} z_{ik}$ 表示是否选择第 $k$ 个用户在任何子载波 (背包) 中进行调度, 并且新的背包容量 (数据时间长度) $W' = BW$; 是由联合背包的总容量给出的.

(2) 导出下界: 下界是将联合背包后选择的用户分成 $B$ 个单独的背包. 这是通过求解一系列子集和问题来实现的, 如下所示:

$$\max_{z} \sum_{k=1}^{N'_k} w_k \cdot z_k \tag{23}$$

$$\text{s.t. } \sum_{k=1}^{N} w_k \cdot z_k \leqslant W_j, \tag{24}$$

$$z_j \in \{0, 1\},\ \text{for } j = 1, 2, \ldots, N'_j. \tag{25}$$

可以注意到, 当前的问题试图在每个单独的背包中适配尽可能多的物品, 其中 $N'_j$ 表示子集中的用户数量, 该数值来自问题 (20)∼(22) 的最优解 $z'$.

(3) 分支定界 (branch-and-bound, BNB) 算法: 该算法是一种基于树搜索的优化技术, 通过计算下界来指导搜索过程. 在决策树的每个节点, 如果通过下界计算发现所有选择的项目 (在计算上界后) 能

够完全装入 $B$ 个背包, 并且下界等于上界, 那么我们可以立即停止搜索并返回当前解, 因为已经找到了最优解. 否则, 如果只能得到部分项目的可行解, 算法将继续探索其他树节点, 考虑包含更多项目的情况, 以寻找更优的解. 在确认了用分支定界算法解决基础数据的放置问题之后, 我们进一步的使用简化的 BNB 方法来处理 AI 数据的放置. 类比基础数据的放置方式, 第一步筛选出当前容量所能容纳的价值最大的用户组 (在 AI 用户稀少的条件下这一步可以忽略), 然后我们按照业务匹配性的高低给出所有用户期望的卫星顺序. 然后逐个将 AI 数据放置在最期望的卫星上, 如果某颗卫星容量已满, 则下一个用户将会避开这个卫星选择列表中期望次之的卫星.

综上所述, 我们将所提出方案描述为 DBNB (双层分支定界) 方案, 为了详细描述这一方案在卫星系统中发挥作用的方式, 我们引入中心卫星的概念, 中心卫星能够起到中心化资源分配的作用. 具体来说主要有以下几个步骤:

(1) 卫星和用户参照 NB-IoT 建立初步联系, 卫星获知地面用户的规模数量以及所要上传的数据包大小;

(2) 主卫星使用 DBNB 算法分配用户数据的上传卫星与所占据信道;

(3) 主卫星将这一分配结果告知辖区内各卫星;

(4) 卫星群将分配结果广播至所覆盖区域, 使用户知晓;

(5) 进行数据收发与在轨计算, 地面用户反馈效果.

多卫星协同场景下的复杂度分析如下: 在系统包含 $J$ 颗卫星、$I$ 个用户的情况下, 分支定界法的时间复杂度主要由两层搜索决定. 基础通信数据分配阶段, 每层分支需求解 $O(N)$ 个 0-1 背包问题, 复杂度为 $O(J \cdot I \cdot W)$; AI 数据放置阶段, 基于业务匹配度排序的复杂度为 $O(I \log I)$, 卫星容量冲突处理的复杂度为 $O(I \cdot J)$. 因此, 整体复杂度为 $O(J \cdot I \cdot W + I \log I)$. 当 $J$ 固定时, 复杂度随 $J$ 呈线性增长, 适用于大规模卫星网络场景. 仿真结果也表明, 当用户数从 100 增至 1000 时, DBNB 方案的方案分配时间增长速率逐渐下降, 验证了其良好的扩展性.

## 5　结果及分析

本节中我们对所提出的方案进行了仿真验证, 考虑用户数量 $(100, 500)$, 用户产生的数据包大小服从 $U(10, 20)$ 的均匀分布, 帧大小为 5000 个时频资源格, AI 用户产生的数据包数量在 $U(20, 40)$ 之间变化, AI 数据包大小在 $U(30, 40)$ 之间变化, 且共有 6 种 AI 业务. 卫星共有 6 颗, 对应着 6 种不同的业务. 数据包大小和计算量并不直接相关, 我们对在轨推理 AI 数据的计算量参数进行归一化处理, 卫星的在轨计算能力 $C_{sat}$ 服从 $U(0.2, 0.6)$ 的均匀分布, 用户的 AI 数据包计算量 $C_{ai}$ 服从 $U(0.5, 1)$ 的均匀分布, 用于分析业务放置问题的计算时间, 用户以 5% 的比例产生 AI 数据包. 信道条件设置方面, 为准确刻画低轨卫星与地面用户链路的特性, 采用以自由空间传播模型计算路径损耗, 作为大尺度平均信号强度的基础; 并引入标准差为 8 dB 的对数正态分布阴影衰落, 以模拟由于地形起伏、建筑物遮挡等环境因素造成的随机信号波动. 为了方便说明, 我们给出本文仿真参数, 如表 2 所示.

对照算法有 3 种, 分别是以数据块重量作为分配目的的 Greedy-Weight 方案、以数据块价值作为分配目的的 Greedy-Value 方案和随机分配的方案. 需要指出的是, Greedy-Value 方案没有分支定界的过程, 仅能够对用户进行初筛. 图 3 给出的是总价值随用户总数量的变化. 从图上可以看出我们所提出的 DBNB 算法具有最好的性能, 要超过 Greedy-Value 和随机的方案. 我们的方案能够对用户的性价比进行排序, 进而得出接近最大价值的方案. 在对照算法中, Greedy-Weight 表现最差, 因为这种方法优先考虑占据点位多的用户, 而非价值高的用户, 因此表现最差. 此外, 随着用户的增多, 低性价比的用户数量也在变多, 这进一步放大了 Greedy-Weight 方案的劣势. 图 4 给出的是接入用户数量随用户总数量的变化. 从图上可以看出我们所提出的 DBNB 算法仍然具有最好的性能, 要超过 Greedy-Value 和随机的方案. 由于我们的方案具备排序的功能, 所以占点位少, 高价值的用户会优先被考虑. 在对照

表 2　仿真参数.

**Table 2**　Simulation parameters.

| Parameter | Value |
|---|---|
| Number of users $N$ | 100, 500 (units) |
| Conventional data packet size $w_i \times b_i$ | $U(10, 20)$ (resource grids) |
| Frame size $W$ | 5000 (resource grids) |
| AI user ratio $\rho_{ai}$ | 5% |
| Number of AI data packets $N_{ai}$ | $U(20, 40)$ (units) |
| AI data packet size $w_i^{ai} \times b_i^{ai}$ | $U(30, 40)$ (resource grids) |
| Number of AI service types $k$ | 6 (types) |
| Number of satellites $K$ | 6 (units) |
| Satellite computing capability $C_{sat}$ | $U(0.2, 0.6)$ (normalized value) |
| AI data packet computation load $C_{ai}$ | $U(0.5, 1)$ (normalized value) |
| Shadow fading standard deviation $\sigma_{shadow}$ | 8 (dB) |

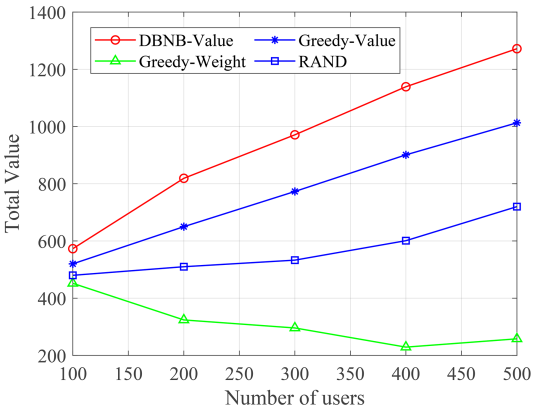


图 3　(网络版彩图) 总价值随用户总数量的变化.

**Figure 3**　(Color online) Total value via user count.

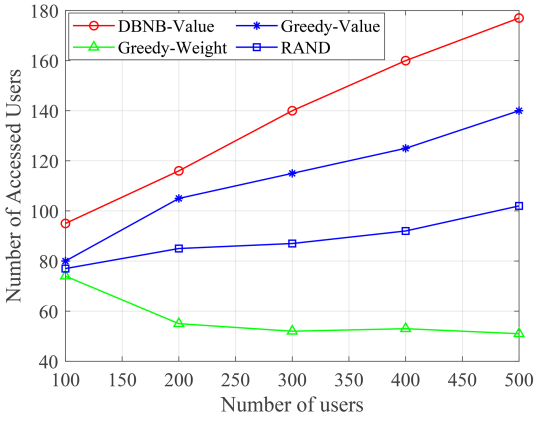


图 4　(网络版彩图) 接入用户数量随用户总数量的变化.

**Figure 4**　(Color online) Access number via user count.

算法中, Greedy-Weight 表现最差, 随着用户的增多, 大体积用户挤占小体积用户的空间, 导致用户总数越多, 接入用户的数量反而越少.

图 5 给出的是方案运行时间随着用户总数量的变化, 随着用户的数量从个位数增加至 1000, 我们所提出的方案 DBNB 耗费了最长的时间, 但是随着用户数量的增多, 时间增长的效果逐步下降. 这是因为随着用户增多, 高性价比用户也在增多, 但是帧的容量有限, 这在一定程度上降低了我们方案的复杂性. 对照算法 Greedy-Value 和随机方案均呈现出极低的复杂度, 但是这两种方法均不能实现较好的数据分配效果.

为验证方案对海量用户的支持能力, 进一步分析用户规模与 AI 在轨推理时间的关系, 图 6 给出的是 AI 在轨平均推理时间随 AI 数据包数量的变化. 当用户规模从 250 增至 500 时, 按照 8% 比例生成的 AI 数据包数量相应从 20 增至 40, 如图 6 所示, DBNB 方案的平均推理时间从 0.05 s 缓慢增至 0.15 s, 而 Greedy-Weight 和随机方案的推理时间增幅则更加显著. 这表明随着用户规模扩大, DBNB 方案通过优先匹配卫星业务类型与 AI 数据需求, 有效控制了推理时间的增长速率, 验证了其在海量接入场景下的稳定性. 在硬件限制方面, 实际星载计算能力已具备工程可行性, 如三体计算星座单星浮点计算能力 744 TOPS, 完全有能力在极短时间内完成 500 用户的 DBNB 算法迭代. 然而, 当用户

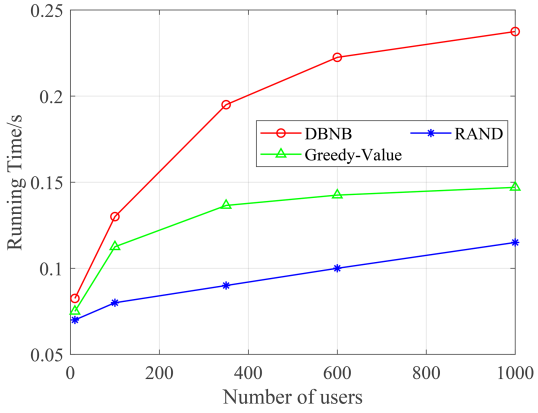


**图 5**　(网络版彩图) 方案分配时间随用户总数量的变化.

**Figure 5**　(Color online) Allocation time via user count.

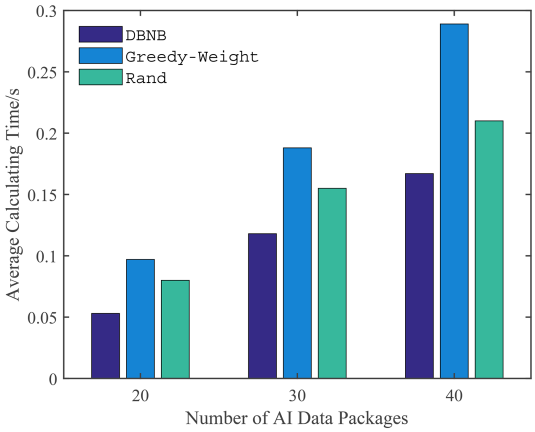


**图 6**　(网络版彩图) 平均 AI 数据的计算时间.

**Figure 6**　(Color online) Average AI calculate time.

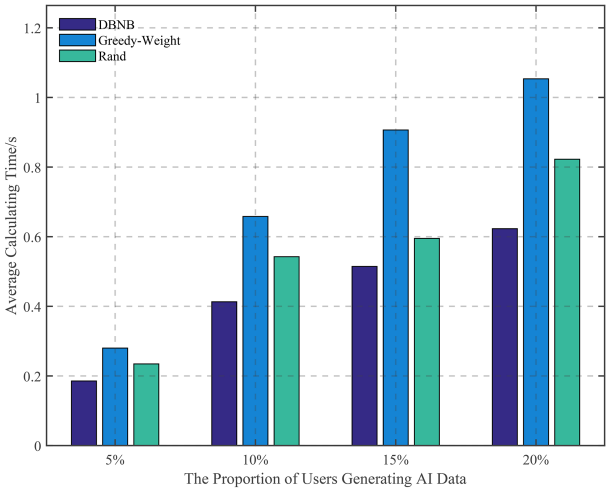


**图 7**　(网络版彩图) AI 数据处理时间随产生 AI 数据的用户比例的变化 (用户数量 = 500).

**Figure 7**　(Color online) The time for AI data processing varies with the proportion of users who generate the AI data (number of users = 500).

规模过大时, 仅靠当前仿真场景设置难以准确描述迭代时间, 可能影响实时性, 我们将在后续的工作中详细探讨这一问题的影响. 为了进一步描述在不同 AI 数据比例下的分配效率, 我们还设置了一组实验, 令 AI 数据和常规数据比例分别为 5%, 10%, 15%, 20%, 并观察他们的效果.

实验通过设置 AI 数据比例从 5%~20% 的梯度变化, 验证了方案在不同数据构成下的鲁棒性, 如图 7 所示. 结果表明, 当 AI 数据比例提升时, DBNB 方案的 A 数据处理时间仍然低于其他对照方法. 标明即使 AI 数据占比增加, 系统仍能通过动态调整帧结构资源块的放置策略, 确保 AI 数据的优先调度. 且方案对数据构成的动态变化表现出良好适应性, 为实际场景中混合业务的高效处理提供了理论支撑.

## 6　结论

本文针对具备智能在轨服务功能的 NTN-IoT 网络, 提出了一种基于帧结构资源部署的优化方案. 该方案面向多颗具备在轨推理功能的卫星服务海量地面用户的场景, 重点解决常规通信数据与 AI 推

理数据的帧结构资源分配问题. 本文将问题建模为 0-1 背包问题, 并创新性地设计了基于分支定界法的求解方法, 该方法优先处理用户通信信息, 再分配 AI 数据, 有效提升了资源利用效率. 仿真结果表明, 所提方案能够显著提升频谱效率, 为海量用户接入提供了有力支持, 同时实现了 AI 数据的合理放置, 降低了平均推理时间. 本研究为未来 NTN-IoT 网络中实现高效、智能的资源管理提供了新的思路和方法.

## 参考文献


1 3rd Generation Partnership Project. Technical Specification Group Radio Access Network; Solutions for NR to support non-terrestrial networks (NTN) (Release 16). 3GPP TR 38.821. 2018
2 Chen L, Yu S, Chen Q, et al. 5S: design and in-orbit demonstration of a multifunctional integrated satellite-based Internet of Things payload. IEEE Int Things J, 2024, 11: 12864–12873
3 Alvarez G, Fraire J A, Hassan K A, et al. Uplink transmission policies for LoRa-based direct-to-satellite IoT. IEEE Access, 2022, 10: 72687–72701
4 Kodheli O, Maturo N, Chatzinotas S, et al. NB-IoT via LEO satellites: an efficient resource allocation strategy for uplink data transmission. IEEE Int Things J, 2022, 9: 5094–5107
5 DeepSeek. DeepSeek artificial intelligence system. https://www.deepseek.com
6 Saifaldawla A, Ortiz F, Lagunas E, et al. GenAI-based models for NGSO satellites interference detection. Trans Mach Learn Comm Netw, 2024, 2: 904–924
7 Li Y, Wang M, Hwang K, et al. LEO satellite constellation for global-scale remote sensing with on-orbit cloud AI computing. IEEE J Sel Top Appl Earth Obs Remote Sens, 2023, 16: 9369–9381
8 Zhang R, Du H, Liu Y, et al. Generative AI agents with large language model for satellite networks via a mixture of experts transmission. IEEE J Sel Areas Commun, 2024, 42: 3581–3596
9 Oligeri G, Sciancalepore S, Raponi S, et al. PAST-AI: physical-layer authentication of satellite transmitters via deep learning. IEEE Trans Inform Forensic Secur, 2023, 18: 274–289


# Uplink frame resource placement scheme design for NTN-IoT with intelligent on-orbit service function


Yu WANG[1], Zhiqiang LI[1], Shuai HAN[1*], Abderrahim BENSLIMANE[2] & Cheng LI[3]

1. *School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China*
2. *Laboratoire Informatique d'Avignon, University of Avignon, Avignon 84911, France*
3. *School of Engineering Science, Simon Fraser University, Burnaby V5A, Canada*
* Corresponding author. E-mail: hanshuai@hit.edu.cn



**Abstract** In this paper, a resource placement optimization scheme based on frame structure is designed for the NTN-IoT network with intelligent on-orbit service functions. Multiple satellites with on-orbit reasoning function serve many ground users, and these ground users mainly produce conventional communication data and a small amount of AI reasoning data. Therefore, satellites need to reasonably arrange the frame structure of these data. In this paper, we formulate the problem according to the frame structure characteristics given by NPUSCH (narrow band uplink shared channel), and further formulate the problem as 0-1 knapsack problem. For the purpose of resource placement efficiency, this paper designs a branch and bound based method, which first processes the communication information of users and then allocates AI data. The simulation results show that the proposed scheme can improve the spectrum efficiency, which provides an advantage for the access of massive users. At the same time, the proposed scheme can also reasonably place the AI data to reduce the average inference time.



**Keywords** NPUSCH, IoT, frame structure, mass access, on-orbit reasoning