Stochastic Delay Analysis for Satellite Data Relay Networks With Heterogeneous Traffic and Transmission Links

Yan Zhu, Di Zhou, *Member, IEEE*, Min Sheng, *Senior Member, IEEE*, Jiandong Li, *Senior Member, IEEE*, and Zhu Han, *Fellow, IEEE*

***Abstract*— The satellite data relay networks (SDRNs) hold great promise in 6G communications for the timely offloading of the global traffic. Since the delay performance is regarded as one of the most important metrics reflecting the offloading efficiency, studying its relationship with network parameters becomes really essential to the development and application of the SDRN. However, the complex data offloading process and heterogeneity of traffic arrivals and transmission links pose many challenges to the stochastic delay analysis. To accurately model the data offloading process in SDRNs, we build a series-parallel queuing model with through and cross traffic while considering the propagation delay. On this basis, we respectively propose a propagation delay embedded min-plus convolution method based on stochastic network calculus and a Markov chain method based on Monte Carlo to depict the leftover services of the heterogeneous links received by the per-flow traffic in an aggregate. To eliminate the impacts of the heterogeneity, we uniformly characterize the arrivals and leftover services by their moment generating functions (MGFs) which contain the full moment information, and shield the heterogeneity by deriving the envelopes of the arrivals and leftover services with the help of MGFs, Chernoff bound and union bound. Then, in the light of the geometric relationship between the envelopes of the arrivals and leftover services, we analyze the upper bounds of the stochastic delay, which provides the guidance to the network configuration. Eventually, simulation results verify the effectiveness of the theoretical analysis and further reveal maximum four times the delay difference between the heterogeneous links influenced by traffic type, burstiness, and access number.**

***Index Terms*— Stochastic delay analysis, heterogeneous traffic and transmission links, series-parallel queuing process, leftover service, stochastic network calculus, Markov chain Monte Carlo.**

Manuscript received March 16, 2020; revised July 2, 2020 and September 1, 2020; accepted September 7, 2020. Date of publication September 21, 2020; date of current version January 8, 2021. This work was supported in part by the Natural Science Foundation of China under Grant U19B2025, Grant 61725103, Grant 61701363, Grant 61931005, and Grant 62001347; and in part by NSF EARS-1839818, CNS1717454, CNS-1731424, and CNS-1702850. The associate editor coordinating the review of this article and approving it for publication was L. X. Cai. *(Corresponding author: Di Zhou.)*

Yan Zhu, Di Zhou, Min Sheng, and Jiandong Li are with the State Key Laboratory of ISN, Institute of Information Science, Xidian University, Xi'an 710071, China (e-mail: yanzhu@stu.xidian.edu.cn; zhoudi@xidian.edu.cn; msheng@mail.xidian.edu.cn; jdli@mail.xidian.edu.cn).

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea (e-mail: zhan2@uh.edu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2020.3023857

I. Introduction

BY VIRTUE of the high coverage of the earth's surface, the satellite data relay network (SDRN) has promising development prospects in 6G communications to compensate for the lack of communication capabilities of terrestrial cellular networks in remote areas [1], [2]. Besides the extensive coverage, plenty of tasks put forward higher requirements for network delay performance. For example, in the emergency rescues of Malaysia Airlines 370 incident and Wenchuan earthquake, the establishment of faster emergency communications means that more lives are possible to be saved. Therefore, studying the relationship between the delay performance of the SDRN and network parameters becomes really essential to guide the resource allocation of 6G communications and better complete communication tasks [3]–[7].

As a typical network structure in space, the SDRN is mainly composed of user satellites (USs) in low earth orbits (LEOs), data relay satellites (DRS) in the geostationary orbit (GEO) and ground stations (GSs). USs have various types and different functions, such as communication satellites, earth observation satellites, meteorological satellites, navigation satellites, etc. As a result, the acquired traffic data must have obvious heterogeneity. Together with the heterogeneity caused by the different offloading processes, they both exert large impacts on the delay performance. The heterogeneity of the traffic data is mainly reflected in the aspects of the arrival rate, burstiness, arrival interval distribution and so on. The heterogeneity of offloading processes is reflected in: (1) direct offloading to the GS through the satellite-ground link (SGL), which is an intermittent offloading process due to the earth occlusion; (2) indirect offloading to the GS through the DRS, which is a two-level offloading process. Heterogeneous traffic data results in differentiated delay characteristics, which will be further highlighted in heterogeneous offloading processes. Hence, investigating the delay performance of the heterogeneous traffic in heterogeneous offloading processes plays an important role in the development and application of the SDRN [8]–[11].

It is worth noting that two challenges are inevitable to be faced during studying the delay performance of the heterogeneous traffic. The first thing that needs to be solved is how to accurately model the heterogeneous offloading processes and depict the leftover transmission services received by the

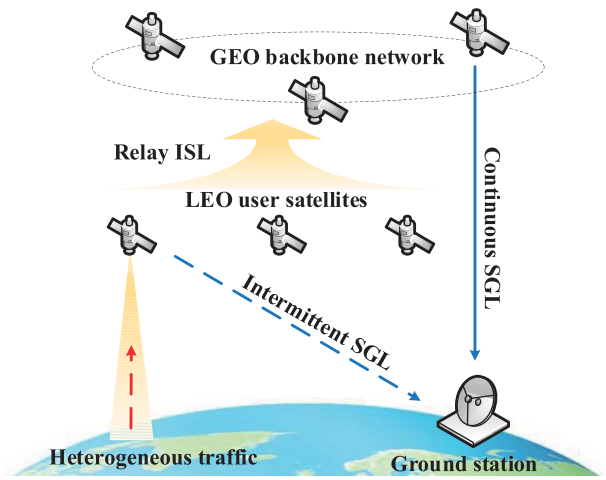

Fig. 1. The common data offloading scenario of SDRNs.

heterogeneous traffic. In view of the cascade and discontinuity exhibited in direct and indirect offloading processes respectively, their transmission behaviors are totally different from each other. Besides, heterogeneous traffic owns different arrival features and receives differentiated transmission services so that it is necessary to accurately model the heterogeneous offloading processes and differentially describe the leftover services received by the heterogeneous traffic. Secondly, what needs to be solved is how to eliminate the impacts of the heterogeneity from the traffic and leftover transmission services on the stochastic delay analysis. Since the stochastic delay is defined as the time difference between traffic arrivals and transmission services, the random distributions of heterogeneous arrivals and services make it impossible to directly perform algebraic calculations on them. Therefore, a unified characterization method is necessary to be proposed to eliminate the impacts of the heterogeneity from the traffic and leftover transmission services on the stochastic delay analysis. Only by overcoming the above challenges, can we make the model more realistic and analyze the stochastic delay performance in SDRNs.

In this paper, aiming at the above challenges, we investigate the stochastic delay analysis in SDRNs integrated by heterogeneous traffic and transmission links, as well as US multiple access. As shown in Fig. 1, our scenario is constituted by USs deployed in LEOs, a DRS backbone network in the GEO, and multiple GSs. The heterogeneous traffic is collected by USs and either offloaded directly to GSs when the SGL is active or forwarded to the DRS backbone network temporarily and finally offloaded to GSs. The main contributions of our work can be summarized as follows:

- To accurately model the heterogeneous offloading processes in SDRNs, we build a series-parallel queuing model with through and cross traffic while considering the propagation delay. On this basis, we respectively propose a propagation delay embedded min-plus convolution method based on stochastic network calculus [12]–[15] and a Markov chain method based on Monte Carlo (MCMC) to depict the leftover services of the heterogeneous offloading processes received by the per-flow traffic in a heterogeneous aggregate.
- To eliminate the impacts of the heterogeneity, we uniformly characterize the arrivals and leftover services by their moment generating functions (MGFs) which contain the full moment information, and shield the heterogeneity by deriving the envelopes of arrivals and leftover services with the help of MGFs, Chernoff bound and union bound. Then, we derive the upper bounds of the stochastic delay in the light of the geometric relationship between the envelopes of the arrivals and leftover services, which provides the guidance to the network configuration according to the delay requirement of the traffic data.

The remainder of the paper is organized as follows: we overview related works in Section II. Then, we elaborate the system model in Section III. Next, we analyze the stochastic delay performance considering the heterogeneous traffic and transmission links as well as US multiple access in Section IV. In Section V, we conduct numerous simulations and evaluations. Finally, we conclude the paper in Section VI.

II. Related Works

Network modeling and performance analysis have been long-term attractive topics. Correspondingly, the related research has also been carried out in full swing.

For the modeling of satellite networks, the mainstream modeling methods can be summarized as the mesh grid [16]–[18], time-expanded graph (TEG) [19]–[21], storage time aggregated graph (STAG) [22], [23], and queuing model [24]–[30]. In [16]–[18], the authors introduce a mesh grid to model the typical satellite constellations, such as the Iridium network and Globalstar. However, not all the satellites are networked in regular constellations, especially for those USs with heterogeneous functions. As a result, the mesh grid has its own boundedness in modeling the irregular structures of multi-layered satellite scenarios. Some studies [19]–[21] involve the TEG to transform the mission maximization problem to a graph-based mixed-integer nonlinear programming. But the missions and transmission are all preset and fixed so that the relationships among the network stochastic performance, traffic features and transmission modes are hardly to be known. The authors in [22], [23] utilize a STAG to depict the dynamic characteristics of the multi-layered satellite network (MLSN) and investigate a graph-based maximum flow problem. Unfortunately, such a kind of the modeling method cannot reflect the heterogeneity and randomness of satellite traffic and transmission links. The existing works based on queuing theory [24]–[30] focus on the system modeling of some relatively simple satellite networks (i.e., with homogeneous traffic and link). The largest resistance of applying the queuing theory to the relatively complex networks is that the limitations of some strong independent assumptions need to be first guaranteed before analysis, which makes it unsuitable to the more general heterogeneous scenarios.

For the characterization of heterogeneous traffic and links, the authors in [31] analyze the backlog bound of a tandem queuing model for the multi-hop vehicular ad hoc network with considering the delay sensitive and tolerant traffic.

However, the tandem transmission among the vehicles is assumed to be homogeneous for the analytical simplicity. [32] studies the optimal task allocation based on martingale theory in heterogeneous vehicular networks jointly considering the service from the cloud center and from the fog node. Its considered traffic is only the homogeneous Markov modulated on-off (MMOO) traffic so that the related model cannot be extended to the heterogeneous condition. Reference [33] proposes a hierarchical coordinated planning architecture to address the issue of the coordinated planning of the heterogeneous earth observation resources including the satellite, airship, and unmanned aerial vehicle. But it does not take the heterogeneous links into account neither. A scheduling strategy for the multimedia is proposed to optimize the average end-to-end delay in high-speed train networks including the heterogeneous service from track-side access points (TAPs) and base stations (BSs) [34]. Although the heterogeneous traffic and links seem to be both involved, due to the time-division switching mechanism, the essence of the link transmission is actually homogeneous. In [35], the service provisioning and user association are investigated for heterogeneous wireless railway networks where the TAP service and BS service coexist. Nevertheless, the related services encounter the similar condition with that in [34].

To sum up, existing works either ignore the inherent characteristics of SDRNs or have no regard for the effect of heterogeneous traffic and links in their system modeling and performance analysis. In this paper, we attempt to overcome the above problems to evaluate the stochastic delay performance of SDRNs.

III. System Model

A. Network Model

Intuitively, we transform the real data offloading scenario to a series-parallel queuing model as shown in Fig. 2. The storages and transmitters of the US and DRS are intuitively modelled as queues and servers, respectively. The tagged US holds two queues for US-DRS-GS link and US-GS link, respectively, which form the serial queuing process and parallel queuing process, correspondingly. Similarly, the DRS holds a queue for the data not transmitted in time. Each pair of serial queues contains the through traffic from tagged USs and the cross traffic from other multiplexed USs. Assume that the scheduling discipline of the heterogeneous traffic is first in first out (FIFO). Because the queuing system can quickly reach the steady state so that the handover of the US among DRSs every few hours makes no difference to the system performance and is reasonably ignored. Moreover, the handover is generally conducted by GS so that the communication delay between DRSs is neither considered [36]–[38].

For the space traffic, we consider two kinds of typical traffic (i.e., Poisson traffic and MMOO traffic) without loss of generality and for the sake of simplicity, which are widely applied to depict the voice, VoIP, image, etc [39]. Note that our model still holds for multiple types and numbers of traffic inputs with the prerequisite that their MGFs exist. From the perspective of the tagged US, part of its collected

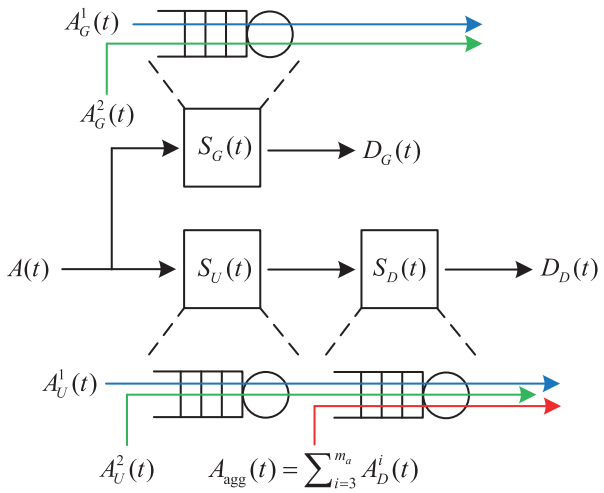

Fig. 2. The series-parallel queuing model. For the tagged US, part of traffic (i.e., Poisson traffic $A_U^1(t)$ and MMOO traffic $A_U^2(t)$) is distributed to the serial queues, which refers to the through traffic. The rest (i.e., Poisson traffic $A_G^1(t)$ and MMOO traffic $A_G^2(t)$) is distributed to the parallel queue. $A_{\text{agg}}(t)$ is the aggregated traffic from other multiplexed USs to DRS queue, which refers to the cross traffic. $S_U(t)$, $S_D(t)$, and $S_G(t)$ indicate the transmission services of the US-DRS link, DRS-GS link, and US-GS link, respectively.

heterogeneous data (i.e., Poisson traffic $A_U^1(t)$ and MMOO traffic $A_U^2(t)$) is offloaded by US-DRS-GS link with probability η. The rest of the data (i.e., Poisson traffic $A_G^1(t)$ and MMOO traffic $A_G^2(t)$) is offloaded by US-GS link with probability $1-\eta$. Note that the split traffic flows still keep their features (i.e., Poisson and MMOO) [40]. $A(t)$ indicates the overall arrival process of a US. $A_D^i(t)$ is the arrival process of the DRS from US i, $i \in \{3, 4, \cdots, m_a\}$ where $m_a - 2$ is the number of other multiplexed USs. $A_{\text{agg}}(t)$ indicates the aggregate arrival process from other multiplexed USs. $S_U(t)$, $S_G(t)$, and $S_D(t)$ represent the service processes of the US-DRS link, US-GS link, and DRS-GS link, respectively. $D_D(t)$ and $D_G(t)$ mean the departure processes of DRS queue and US queue to US-GS link, respectively. The main notations throughout this paper are listed in Table I.

B. Traffic Model

1) Poisson Traffic: Let $N_U^1(t)$ and $N_G^1(t)$ denote the number of newly arrived Poisson packets at the serial and parallel queues in duration $[0, t]$, respectively. If the arrival intervals of these packets obey exponential distribution, the counting process $N_U^1(t)$ is said to be the Poisson process where the probability density function can be written as

$$P(N_U^1(t) = k) = \frac{e^{-\eta\lambda t}(\eta\lambda t)^k}{k!}, \tag{1}$$

where $\eta\lambda$ represents the arrival rate. The cumulative arrival process can be represented as

$$A_U^1(t) = N_U^1(t)/v, \tag{2}$$

where $1/v$ denotes the packet size. Correspondingly, $A_G^1(t)$ can be similarly expressed by changing $\eta\lambda$ to $(1-\eta)\lambda$.

TABLE I
MAIN NOTATIONS

Symbol	Description	Symbol	Description
$A(t)$	overall traffic arrival process of a US	$A_U^1(t)$	Poisson traffic distributed to serial queues
$A_U^2(t)$	MMOO traffic distributed to serial queues	$A_G^1(t)$	Poisson traffic distributed to parallel queue
$A_G^2(t)$	MMOO traffic distributed to parallel queue	$A_{\text{agg}}(t)$	aggregate traffic to DRS queue
$A_D^i(t)$	traffic arrival from other multiplexed USs	$S_U(t)$	service process of US-DRS link
$S_G(t)$	service process of US-GS link	$S_D(t)$	service process of DRS-GS link
$D_D(t)$	departure process of DRS queue	$D_G(t)$	departure process of US queue to US-GS link
η	traffic distribution ratio to serial queues	λ	overall arrival rate of Poisson traffic
$1/v$	packet size	$\boldsymbol{T_a}$	transition probability matrix of MMOO traffic
p_{ij}^a	arrival transition probability from states i to j	M	dimension of $\boldsymbol{T_a}$
$\boldsymbol{B}$	arrival rate vector of MMOO traffic	B_i	arrival rate in state i of MMOO traffic
$\boldsymbol{\pi_a}$	steady-state probability of $\boldsymbol{T_a}$	SNR_{mn}	SNR between two facilities (m,n)
h_{mn}	channel fading coefficient	S_{mn}	slant range between two facilities
γ	path loss exponent	$P_{t_{mn}}$	transmitting power from m to n
N_0	additive Gaussian noise power	K	power ratio of direct path to scattered paths
ω	total power of LOS and scattering signals	$I_0(\cdot)$	first kind of zero order modified Bessel function
$C_{mn}(t)$	instantaneous channel capacity between (m,n)	B_{mn}	channel bandwidth between (m,n)
$C_{UD}(t)$	channel capacity between US and DRS	$A_U(t)$	arrival process of US queue to serial queues
$D_U(t)$	departure process of US queue to serial queues	$S_U(t)$	service process of US queue to serial queues
d_{UD}	propagation delays between US and DRS	$S_{UD}(t)$	slant range between US and DRS
c	propagation rate of the electromagnetic wave	$A_D(t)$	arrival process of DRS queue
$D_D(t)$	departure process of DRS queue	$S_D(t)$	service process of DRS-GS link
$S_{\text{lo}}^{UD1}(t)$	leftover service from serial service for $A_U^1(t)$	$S_{\text{lo}}^{UD2}(t)$	leftover service from serial service for $A_U^2(t)$
$S_{\text{lo}}^{D12}(t)$	leftover service from DRS for $A_U^1(t)$ and $A_U^2(t)$	$S_{\text{lo}}^{G1}(t)$	leftover service from US-GS link for $A_G^1(t)$
$\rho_{A_G^1}$	slope of arrival envelope of $A_G^1(t)$	$N+1$	number of active and inactive periods
S_i	number of inactive periods	S_a	number of active periods
$\boldsymbol{p^r}$	proposal transition matrix	ζ	acceptance distribution
$\boldsymbol{p^s}$	service transition probability matrix	$\boldsymbol{T_s}$	state transition probability matrix
$\mathbb{E}_{X_j}$	expectation of (in)active period samples X_j	$\mathbb{V}_{X_j}$	variance of (in)active period samples X_j
$C_{UG}(t)$	channel capacity between US and GS	θ	free parameter
$M_{A_U^1}(\theta,t)$	moment generating function of $A_U^1(t)$	$\epsilon_{A_U^1}$	violation probability of $A_U^1(t)$
w_{cas}	delay bound of cascaded links	w_{e2e}	end-to-end delay bound from US to GS via DRS
d_{DG}	propagation delay between DRS and GS	$\rho_{A_U^1}$	slope of arrival envelope of $A_U^1(t)$
$\rho_{S_{\text{lo}}^{D12}}$	slope of service envelope of $S_{\text{lo}}^{D12}(t)$	$\rho_{A_{\text{agg}}}$	slope of arrival envelope of $A_{\text{agg}}(t)$
ρ_{S_D}	slope of service envelope of $S_D(t)$	$\rho_{S_{\text{lo}}^{UD1}}$	slope of service envelope of $S_{\text{lo}}^{UD1}(t)$
ρ_{S_U}	slope of service envelope of $S_U(t)$	w_{dis}	delay bound of discontinuous link
$\rho_{S_{\text{lo}}^{G1}}$	slope of service envelope of $S_{\text{lo}}^{G1}(t)$	ρ_{S_G}	slope of service envelope of $S_G(t)$

2) Markov Modulated on-off Traffic: A Markov modulated on-off traffic is determined by *i)* the arrival transition probability matrix $\boldsymbol{T_a} = \{p_{ij}^a\}, i,j \in \{1,2,\cdots,M\}$ [28], where M represents the number of the arrival state $X(t)$ that denotes a finite, irreducible, continuous-time Markov chain (CTMC) with state space $S_X = \{1,2,\cdots,M\}$, and *ii)* the arrival rate vector $\boldsymbol{B} = \{B_1, B_2, \cdots, B_M\}$. For instance, the transition matrix of the Markov chain when $M = 2$ is given by

$$\boldsymbol{T_a} = \begin{bmatrix} p_{11}^a & 1-p_{11}^a \\ 1-p_{22}^a & p_{22}^a \end{bmatrix}, \tag{3}$$

where the steady-state probability of the transition probability matrix is given by

$$\boldsymbol{\pi_a} = \left(\frac{1-p_{22}^a}{2-p_{11}^a-p_{22}^a}, \frac{1-p_{11}^a}{2-p_{11}^a-p_{22}^a} \right). \tag{4}$$

The cumulative arrival process of MMOO traffic at the serial queues can be represented as

$$A_U^2(t) = \sum_{\tau=1}^{t} \eta f_{\boldsymbol{B}}(\tau), \tag{5}$$

where $f_{\boldsymbol{B}}(\tau) \in \boldsymbol{B}$. Similarly, $A_G^2(t)$ related to parallel queue can be similarly expressed by changing η to $(1-\eta)$.

3) Aggregated Traffic: The traffic from other multiplexed USs that arrives at the queue of the DRS refers to the cross traffic, which contributes to the statistical multiplexing. The aggregated arrival process from other multiplexed USs is

$$A_{\text{agg}}(t) = \sum_{i=3}^{m_a} A_D^i(t), \tag{6}$$

where $A_D^i(t)$ is the arrival process of US i to the DRS during $(0,t]$. m_a is m_a-th aggregated number. The aggregated traffic can be any form, i.e., heterogeneous or homogeneous.

C. Service Model

1) Channel Model: In satellite networks, the line of sight (LOS) signal is dominant so that the wireless channel for satellite networks can be modelled as a Rician fading channel with additive Gaussian noise [41]–[43]. The signal-to-noise ratio (SNR) between two facilities (m,n) in satellite networks

can be expressed as

$$\mathrm{SNR}_{mn}(t) = \frac{|h_{mn}|^2 S_{mn}^{-\gamma}(t) P_{t_{mn}}}{N_0}, \tag{7}$$

where $S_{mn}(t)$ represents the slant range between two space facilities. γ is the path loss exponent. $P_{t_{mn}}$ denotes the transmitting power. N_0 is the additive Gaussian noise power. h_{mn} indicates the channel fading coefficient that is the function of Ω and K, where Ω indicates the total power of LOS and scattering signals. K denotes the ratio between the power in the direct path and in other scattered paths [41], [44]. On this basis, the relationship among instantaneous channel capacity $C_{mn}(t)$ between two facilities (m, n), $\mathrm{SNR}_{mn}(t)$, and channel bandwidth B_{mn} can be further obtained by the Shannon formula [45].

2) Serial Service: For the tagged US, the serial service can be split into two parts that are the primary service from the US to DRS and the secondary service from the DRS to GS, respectively. In the primary part, the cumulative service is denoted by $S_U(t) = S_U(0,t) = \sum_{\tau=0}^{t} C_{UD}(\tau)$. Meanwhile, let $A_U(t)$ and $D_U(t)$ represent the arrival and departure processes, respectively, corresponding to service process $S_U(t)$ where $A_U(t) = A_U^1(t) + A_U^2(t)$ and $D_U(t) = D_U^1(t) + D_U^2(t)$. $A_U^i(t)$ and $D_U^i(t)$, $i \in \{1, 2\}$, indicate the arrival processes and departure processes of Poisson traffic and MMOO traffic, respectively. For departure process $D_U(t)$, we have

$$D_U(\tau_2) \geq \inf_{0 \leq \tau_1 \leq \tau_2} \{A_U(\tau_1) + S_U(\tau_1, \tau_2)\}. \tag{8}$$

In the secondary part, the cumulative service is denoted by $S_D(t) = S_D(0,t)$. Similarly, let $A_D(t)$ and $D_D(t)$ represent the arrival and departure process corresponding to service process $S_D(t)$. Due to the relatively large propagation delay (on the order of hundreds of milliseconds) from the US to DRS, it cannot be ignored in the end-to-end delay analysis. Hence, the traffic arrival time of the secondary link should be the sum of the traffic departure time from the primary link and the propagation delay. Because the variation fluctuation of the slant range between the US and DRS is really small and we are pursuing the performance upper bound, we have $A_D(t + d_{UD}) = D_U(t)$ where d_{UD} is the propagation delay between the US and DRS. Specifically, $d_{UD} = \hat{S}_{UD}/c$ where $\hat{S}_{UD}$ means the maximum value of $S_{UD}(t)$ and c is the propagation rate of the electromagnetic wave. Correspondingly, we have $A_D^i(t + d_{UD}) = D_U^i(t)$, $i \in \{1, 2\}$. For the departure process $D_D(t)$, we have

$$D_D(d_{UD}, t) \geq \inf_{0 \leq \tau_2 \leq t - d_{UD}} \{A_D(\tau_2 + d_{UD}) + S_D(\tau_2 + d_{UD}, t)\}. \tag{9}$$

We define the concept of leftover service $S_{\mathrm{lo}}^{UD1}(d_{UD}, t)$ in the serial service that means the per-flow service received by traffic $A_U^1(t)$ in the multiplexed traffic flows. Hence, the expression of leftover service $S_{\mathrm{lo}}^{UD1}(d_{UD}, t)$ is shown in Theorem 1.

Theorem 1: The expression of the leftover service for Poisson traffic in the cascaded link can be written as

$$S_{lo}^{UD1}(d_{UD}, t) = [S_U \otimes S_{lo}^{D12}(d_{UD}, t) - A_U^2(t - d_{UD})]_+, \tag{10}$$

where $S_{lo}^{D12}(d_{UD}, t)$ means the leftover service for Poisson traffic and MMOO traffic in the SGL from d_{UD} to t. $S_U \otimes S_{lo}^{D12}(d_{UD}, t) \triangleq \inf_{0 \leq \tau_1 \leq t - d_{UD}} \{S_U(d_{UD}, \tau_1 + d_{UD}) + S_{lo}^{D12}(\tau_1 + d_{UD}, t)\}$, called the min-plus convolution operation. $[\cdot]_+ = \max(0, \cdot)$. Similarly, the expression of the leftover service for MMOO traffic in the cascaded link can be written as

$$S_{lo}^{UD2}(d_{UD}, t) = [S_U \otimes S_{lo}^{D12}(d_{UD}, t) - A_U^1(t - d_{UD})]_+. \tag{11}$$

Proof: See Appendix A. □

3) Parallel Service: Different from the static relative position between the DRS in the GEO and the GS, the relative position between the US in the LEO and the GS changes over time. Due to the block of the earth, the US can only achieve data offloading when it flies over the visible area of the GS. Moreover, the earth rotates all the time so that the duration of the active and inactive periods is also time-varying and does not obey a specific distribution (e.g., exponential distribution). In this circumstance, the classical queuing model with queue state based vacation policies are no longer suitable. In virtue of the good performance of the MCMC in matching the stochastic distribution of arbitrary samples, we are inspired to construct a Markov chain to describe the stochastic distributions of the active and inactive periods. The states of the Markov chain represent the different durations of the active and inactive periods. Meanwhile, the related steady-state probability corresponds to the occurrence probability of each active period or inactive period. As shown in Fig. 3, $N+1$, corresponding to the number of the states in the Markov chain, equals the sum of the number of different active periods S_a and the number of different inactive periods S_i, which can be obtained by making the statistics of the durations (in minutes) and occurrence numbers of active periods and inactive periods from a satellite simulation tool, i.e., satellite tool kit (STK). It can be seen that the state transition can only happens from the active states to the inactive states or vice versa. This kind of design guarantees the alternative appearance of these two kinds of states, which makes the model more realistic and accurate. Refer to [46], we initiate the proposal transition matrix $\boldsymbol{p^r}$ with non-zero elements

$$p^r_{(i<S_a),(j\geq S_a)} = \frac{1}{S_i}, \tag{12}$$

$$p^r_{(i\geq S_a),(j<S_a)} = \frac{1}{S_a}. \tag{13}$$

Then, we define acceptance distribution $\boldsymbol{\zeta}$ as

$$\zeta_{ij} = \begin{cases} 0, & \text{if } p^r_{i,j} = 0, \\ \min\left(1, \dfrac{\pi_{o_j} p^r_{j,i}}{\pi_{o_i} p^r_{i,j}}\right), & \text{otherwise.} \end{cases} \tag{14}$$

Furthermore, we can obtain the expression of the service transition probability among the states of the Markov chain shown in Fig. 3 as

$$\boldsymbol{p^s} = \boldsymbol{p^r} \boldsymbol{\zeta}, \tag{15}$$

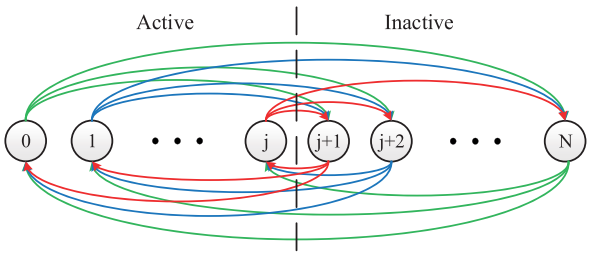

Fig. 3. The state transition of the Markov chain simulates the parallel service from the US to GS.

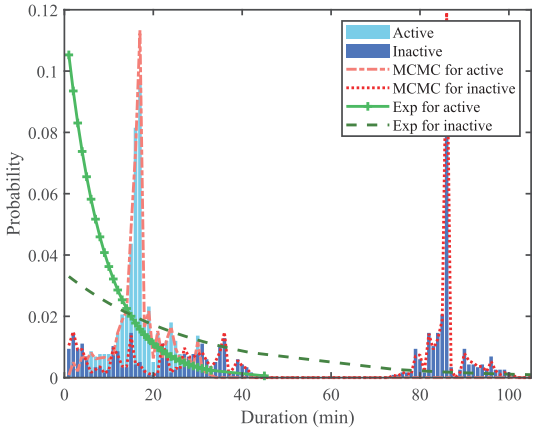

Fig. 4. The statistic distributions of active and inactive periods of the SGL between the US and GS, and the matching performance of the MCMC and the exponential distribution.

where $p^s_{i,i} = 1 - \sum_{j \neq i} p^s_{i,j}$, $i, j \in \{0, 1, \cdots, N\}$. As shown in Fig. 4, it can be seen that the distributions of the durations of active periods and inactive periods are irregular (respectively marked with the light and dark blue bar charts). Since the probability density function of the exponential distribution exhibits a downtrend with the increase of the duration, the matching effect of the MCMC (light and dark red lines marked with "MCMC for active" and "MCMC for inactive" respectively) is much more accurate than that of the simpler two-state on-off Markov chain where the sojourn time in both two states obeys the exponential distribution (light and dark green lines marked with "Exp for active" and "Exp for inactive" respectively).

After solving the problem of the alternative transition between the active and inactive period, the other problem is how to guarantee the sojourn time stayed in each state to emulate the time-varying duration of these two kinds of periods. Herein, we split each state (i.e., active or inactive) into many sub-states. The number and the transition probability of sub-states determine the sojourn time of each state. Due to the memoryless feature of the Markov chain, the sojourn time in each sub-state obeys the exponential distribution. As we know, the stochastic variable constituted by the sum of the exponentially distributed variables obeys the Gamma distribution that could have a pulse shape with duration determined expectation and small enough variance. Hence, we design the unidirectional state transition between sub-states with the state transition probability matrix as

$$\boldsymbol{T_s} = \begin{bmatrix} \boldsymbol{\psi_{00}} & \cdots & \boldsymbol{\psi_{0N}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\psi_{N0}} & \cdots & \boldsymbol{\psi_{NN}} \end{bmatrix}, \tag{16}$$

where the diagonal elements are detailed as

$$\boldsymbol{\psi_{jj}} = \begin{bmatrix} & 1 & & \\ & & 1 & \\ & & & \ddots & \\ & & & & 1 \\ p^s_{j,j} & & & & \end{bmatrix}_{\alpha_j \times \alpha_j}, \tag{17}$$

and the non-diagonal elements are

$$\boldsymbol{\psi_{(j<S_a),(j' \geq S_a)}} = \boldsymbol{\psi_{(j \geq S_a),(j'<S_a)}} = \begin{bmatrix} & \\ & \\ p^s_{j,j'} & \end{bmatrix}_{\alpha_j \times \alpha_{j'}}, \tag{18}$$

where $\alpha_j = \mathbb{E}^2_{X_j} / \mathbb{V}_{X_j}$ according to the feature of the Gamma distribution. $\mathbb{E}_{X_j}$ and $\mathbb{V}_{X_j}$ are the expectation and the variance of the active and inactive period samples X_j, $j \in \{0, 1, \cdots, N\}$, respectively. From this point, we can see that the main differences between the active states and inactive states are mainly reflected in two aspects: 1) the numbers of themselves, i.e. S_a and S_i, 2) the numbers of their sub-states inside, i.e., α_j. The cumulative service process in the parallel queue can be represented as

$$S_G(t) = \sum_{\tau=1}^{t} f_{\boldsymbol{C}}(\tau), \tag{19}$$

where $f_{\boldsymbol{C}}(\tau) \in \boldsymbol{C} = \{\underbrace{C_{UG}(\tau), \cdots, C_{UG}(\tau)}_{\sum_{j=0}^{S_a-1} \alpha_j}, \underbrace{0, \cdots, 0}_{\sum_{j=S_a}^{N} \alpha_j}\}$.

Herein, $\sum_{j=0}^{S_a-1} \alpha_j$ is the total number of the sub-states inside active states and $\sum_{j=S_a}^{N} \alpha_j$ is the total number of the sub-states inside inactive states.

With the similar derivation in Section III-C.2, we have the relationship between the departure process of traffic i ($i \in \{1, 2\}$) and its arrival and service processes as

$$D^i_G(\tau_2) \geq \inf_{0 \leq \tau_1 \leq \tau_2} \{A^i_G(\tau_1) + [S_G(\tau_1, \tau_2) - A^{2-i+1}_G(\tau_1, \tau_2)]_+\}. \tag{20}$$

Intuitively, we introduce the concept of leftover service $S^{Gi}_{\text{lo}}(t)$ separated from the parallel service $S_G(t)$ [12]. $S^{Gi}_{\text{lo}}(t)$ means the per-flow service received by traffic $A^i_G(t)$ in an aggregate $A^1_G(t) + A^2_G(t)$. Therefore, the expression of $S^{Gi}_{\text{lo}}(t)$ can be written as

$$S^{Gi}_{\text{lo}}(t) = [S_G(t) - A^{2-i+1}_G(t)]_+, \tag{21}$$

where $i \in \{1, 2\}$.

IV. Stochastic Delay Analysis

In this section, we analyze the relationship among the link stochastic delay, traffic and link heterogeneity, and US access number.

A. Cascaded Link

1) Arrival and Service: For the derivation of the stochastic delay, we first introduce the concept of the MGF that is an alternative specification of the probability distribution of stochastic variables [12]. For the arrival process of Poisson traffic to the tagged US, we define its MGF during $[\tau, t]$ as

$$M_{A_U^1}(\theta, t-\tau) \triangleq \mathbb{E}\{e^{\theta A_U^1(\tau,t)}\}. \tag{22}$$

Moreover, we define the related MGF envelope as

$$M_{A_U^1}(\theta, t-\tau) \leq e^{\theta(\rho_{A_U^1}(t-\tau)+\sigma_{A_U^1})}, \tag{23}$$

where $\rho_{A_U^1} > 0$ and $\sigma_{A_U^1} \geq 0$. Then, for $\forall \tau \in [0, t]$, we have the related exponentially bounded burstiness [12] as

$$P\{A_U^1(\tau,t) > \rho_{A_U^1}(t-\tau) + b_{A_U^1}\} \leq \epsilon_{A_U^1}(b_{A_U^1}). \tag{24}$$

By introducing a violation probability, the deterministic envelope is relaxed to the stochastic version that makes the envelope tighter. The expression of the violation probability is defined as

$$\epsilon_{A_U^1}(b_{A_U^1}) = \alpha e^{-\theta b_{A_U^1}}, \tag{25}$$

where $b \geq 0$ and $\alpha \geq 0$. With the help of the Chernoff bound that is expressed as

$$P\{X \geq x\} \leq e^{-\theta x}\mathbb{E}\{e^{\theta X}\}, \tag{26}$$

(24) can be further derived as

$$P\{A_U^1(\tau,t) > \rho_{A_U^1}(t-\tau) + b_{A_U^1}\} \leq e^{-\theta(\rho_{A_U^1}(t-\tau)+b_{A_U^1})}\mathbb{E}\{e^{\theta A_U^1(\tau,t)}\}. \tag{27}$$

Combining (27) and (23), we have

$$\begin{aligned} &P\{A_U^1(\tau,t) > \rho_{A_U^1}(t-\tau) + b_{A_U^1}\} \\ &\leq e^{-\theta(\rho_{A_U^1}(t-\tau)+b_{A_U^1})} e^{\theta(\rho_{A_U^1}(t-\tau)+\sigma_{A_U^1})} \\ &= e^{-\theta b_{A_U^1}} e^{\theta\sigma_{A_U^1}}. \end{aligned} \tag{28}$$

Comparing (28) with (25), we have

$$\alpha = e^{\theta\sigma_{A_U^1}}. \tag{29}$$

In order to extend the envelope in (24) corresponding to the arbitrary and fixed τ to a more general form that is suitable to the whole domain of the definition of τ, we have

$$P\{\exists \tau \in [0,t] : A_U^1(\tau,t) > \rho'_{A_U^1}(t-\tau) + b_{A_U^1}\} \leq \epsilon'_{A_U^1}(b_{A_U^1}). \tag{30}$$

With the help of the Union bound that is expressed as

$$P\{\exists i : X_i \geq x\} \leq \sum_i P\{X_i \geq x\}, \tag{31}$$

(30) can be further derived as

$$\begin{aligned} &P\{\exists \tau \in [0,t] : A_U^1(\tau,t) > \rho'_{A_U^1}(t-\tau) + b_{A_U^1}\} \\ &\leq \sum_{\tau=0}^{t} P\{A_U^1(\tau,t) > \rho'_{A_U^1}(t-\tau) + b_{A_U^1}\}. \end{aligned} \tag{32}$$

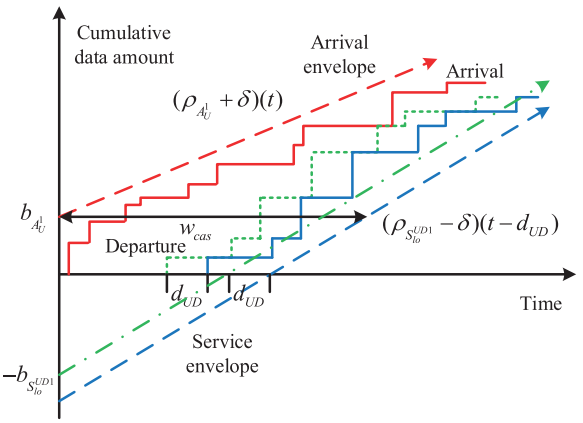

Fig. 5. An illustration of the relationship between the arrival and service envelopes and delay bound.

Let $\rho'_{A_U^1}(t-\tau) = \rho_{A_U^1}(t-\tau) + \delta(t-\tau)$ where $\delta > 0$ is a free parameter and can be optimized. Combining (28) and (32), we have

$$\begin{aligned} &P\{\exists \tau \in [0,t] : A_U^1(\tau,t) > \rho'_{A_U^1}(t-\tau) + b_{A_U^1}\} \\ &\leq \sum_{\tau=0}^{t} P\{A_U^1(\tau,t) > \rho_{A_U^1}(t-\tau) + \delta(t-\tau) + b_{A_U^1}\} \\ &\leq \sum_{\tau=0}^{t} e^{-\theta(b_{A_U^1}+\delta(t-\tau))} e^{\theta\sigma_{A_U^1}} \\ &\leq e^{-\theta b_{A_U^1}} e^{\theta\sigma_{A_U^1}} \sum_{\tau=0}^{\infty} e^{-\theta\delta\tau} = \frac{e^{-\theta b_{A_U^1}} e^{\theta\sigma_{A_U^1}}}{1-e^{-\theta\delta}}. \end{aligned} \tag{33}$$

Recall (30), the expression of $\epsilon'_{A_U^1}(b_{A_U^1})$ is

$$\epsilon'_{A_U^1}(b_{A_U^1}) = \frac{e^{-\theta b_{A_U^1}} e^{\theta\sigma_{A_U^1}}}{1-e^{-\theta\delta}}. \tag{34}$$

Considering Poisson traffic, we have the following theorem about the exponentially bounded fluctuation (EBF) [12] of its serial service during $[d_{UD}, t]$, which is depicted in Theorem 2.

Theorem 2: For the stochastic arrival of Poisson traffic that receives the serial service of S_{lo}^{UD1}, the related EBF of this serial service is

$$\begin{aligned} &P\{\exists \tau \in [d_{UD}, t] : S_{lo}^{UD1}(\tau,t) < \rho'_{S_{lo}^{UD1}}(t-\tau) - b_{S_{lo}^{UD1}}\} \\ &= \frac{e^{-\theta b_{S_{lo}^{UD1}}} e^{\theta(\sigma_{S_U}+\sigma_{S_{lo}^{UD12}}+\sigma_{A_U^2})}}{(1-e^{-\theta\delta})^2} \\ &= \epsilon'_{S_{lo}^{UD1}}(b_{S_{lo}^{UD1}}). \end{aligned} \tag{35}$$

Proof: See Appendix B. □

2) Stochastic Delay: From Fig. 5, the red polyline and the green polyline indicate the actual arrival process and departure process, respectively. In the case of considering the propagation delay d_{UD}, the data have to wait an extra d_{UD} so that the green polyline is moved d_{UD} to the right and becomes the blue polyline. Correspondingly, the service envelope (plotted by green dotted line) is moved d_{UD} to the right and becomes the blue dotted line. As we know,

the instantaneous delay is defined as the horizontal distance between the arrival process and departure process. Therefore, the delay bound is defined as the maximum distance between the arrival envelope and service envelope. For the assurance of the system stability, the slope of the arrival envelope must be less than that of the service envelope, i.e., $\rho_{A_U^1}+\delta < \rho_{S_{\text{lo}}^{UD1}}-\delta$. Correspondingly, the delay bound of the cascaded links can be derived as

$$w_{\text{cas}} = \frac{b_{A_U^1} + d_{UD}(\rho_{S_{\text{lo}}^{UD1}} - \delta) + b_{S_{\text{lo}}^{UD1}}}{\rho_{S_{\text{lo}}^{UD1}} - \delta}, \tag{36}$$

where $b_{A_U^1}$ and $-b_{S_{\text{lo}}^{UD1}}$ indicate the intercepts of the arrival envelope and service envelope, respectively. We choose $\epsilon'_{A_U^1}(b_{A_U^1}) = \epsilon'_{S_{\text{lo}}^{UD1}}(b_{S_{\text{lo}}^{UD1}}) = \epsilon'/2$, from (34), the expression of $b_{A_U^1}$ can be derived as

$$b_{A_U^1} = \sigma_{A_U^1} - \frac{1}{\theta}\left(\ln\left(\frac{\epsilon'}{2}\right) + \ln\left(1 - e^{-\theta\delta}\right)\right). \tag{37}$$

Similarly, from (35), the expression of $b_{S_{\text{lo}}^{UD1}}$ can be derived as

$$b_{S_{\text{lo}}^{UD1}} = \sigma_{S_U} + \sigma_{S_{\text{lo}}^{D12}} + \sigma_{A_U^2} - \frac{1}{\theta}\left(\ln\left(\frac{\epsilon'}{2}\right) + 2\ln\left(1 - e^{-\theta\delta}\right)\right). \tag{38}$$

Because we consider the end-to-end delay that is counted from being collected by the US to arriving at the GS and the propagation delay from the DRS to the GS is on the order of hundreds of milliseconds that cannot be ignored, we finally have the end-to-end delay of the cascaded links as

$$w_{\text{e2e}} = \frac{b_{A_U^1} + d_{UD}(\rho_{S_{\text{lo}}^{UD1}} - \delta) + b_{S_{\text{lo}}^{UD1}}}{\rho_{S_{\text{lo}}^{UD1}} - \delta} + d_{DG}, \tag{39}$$

where d_{DG} is the propagation delay between the DRS and GS. Specifically, $d_{DG} = S_{DG}/c$ where S_{DG} is the slant range between the DRS and GS.

For the Poisson traffic, its probability density function is $P\{N_U^1(t) = k\} = \frac{(\eta\lambda t)^k}{k!}e^{-\eta\lambda t}$. Assume that the packet has a constant size $1/v$, the related MGF is derived as

$$\begin{aligned} M_{A_U^1}(\theta, t) &= \mathbb{E}\{e^{\theta N_U^1(t)/v}\} \\ &= \sum_{k=0}^{\infty} e^{\theta k/v} P\{N_U^1(t) = k\} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} e^{\theta k/v - \eta\lambda t}(\eta\lambda t)^k \\ &= e^{\eta\lambda t(e^{\theta/v}-1)}. \end{aligned} \tag{40}$$

Recall the definition of the MGF envelope, we have

$$\rho_{A_U^1} = \frac{\eta\lambda(e^{\theta/v} - 1)}{\theta}, \tag{41}$$

and $\sigma_{A_U^1} = 0$.

As for the MMOO traffic, the slope of its arrival envelope can be obtained by Theorem 3.

Theorem 3: For the MMOO traffic, the slope of its arrival envelope can be written as

$$\rho_{A_U^2} = \frac{1}{\theta}\ln(sp(\boldsymbol{D_a}(\theta)\boldsymbol{T_a})), \tag{42}$$

where $sp(\cdot)$ represents the spectral radius of the matrix inside and $\boldsymbol{D_a}(\theta)$ is defined as

$$\boldsymbol{D_a}(\theta) = \begin{bmatrix} M_{f_{B_1}}(\theta) & 0 \\ 0 & M_{f_{B_2}}(\theta) \end{bmatrix}. \tag{43}$$

Herein, $M_{f_{B_i}}(\theta) = e^{\theta B_i}$ where $i \in \{1, 2\}$ because we mainly consider two-state MMOO traffic, e.g., image. $\sigma_{A_U^2} = 0$.

Proof: See Appendix C. □

Moreover, the envelope of the leftover service of the secondary service for both Poisson traffic and MMOO traffic can be written as

$$\rho_{S_{\text{lo}}^{D12}} = \rho_{S_D} - \rho_{A_{\text{agg}}}, \tag{44}$$

$$\sigma_{S_{\text{lo}}^{D12}} = \sigma_{S_D} + \sigma_{A_{\text{agg}}}, \tag{45}$$

where $\rho_{A_{\text{agg}}} = \sum_{i=3}^{m_a} \rho_{A_D^i}$ and $\sigma_{A_{\text{agg}}} = \sum_{i=3}^{m_a} \sigma_{A_D^i}$. Recall that satellite links are modelled as Rician channels, we can obtain the slope of envelope ρ_{S_D} of $S_D(t)$ referring to [47], [48] and $\sigma_{S_D} = 0$.

Similarly, the envelope of the leftover service of the serial service for Poisson traffic can be written as

$$\rho_{S_{\text{lo}}^{UD1}} = \min(\rho_{S_U}, \rho_{S_{\text{lo}}^{D12}}) - \rho_{A_U^2}, \tag{46}$$

$$\sigma_{S_{\text{lo}}^{UD1}} = \sigma_{S_U} + \sigma_{S_{\text{lo}}^{D12}} + \sigma_{A_U^2}, \tag{47}$$

where ρ_{S_U} shares the same condition with ρ_{S_D} and $\sigma_{S_U} = 0$. The derivation of the delay bound of MMOO traffic follows the similar procedure that is omitted here to save space.

B. Discontinuous Link

For the US-GS discontinuous link, the data offloading inevitably experiences the intermittent connectivity due to the block of the earth. Herein, we investigate the delay performance of the discontinuous link in this subsection. Compared with the distance of the inter-satellite link (ISL) between the US and DRS, the distance of the SGL between the US and GS is much shorter and the corresponding propagation delay is on the order of 10^{-3} second which is really small so that we rationally ignore the propagation delay of the US-GS link in this subsection.

Following the similar procedure of the arrival envelope derivation in Section IV-A, we have the delay bound as

$$w_{\text{dis}} = \frac{b_{A_G^1} + b_{S_{\text{lo}}^{G1}}}{\rho_{S_{\text{lo}}^{G1}} - \delta}, \tag{48}$$

where $b_{A_G^1}$ is written as

$$b_{A_G^1} = \sigma_{A_G^1} - \frac{1}{\theta}\left(\ln\left(\frac{\epsilon'}{2}\right) + \ln(1 - e^{-\theta\delta})\right), \tag{49}$$

and $b_{S_{\text{lo}}^{G1}}$ can be written as

$$b_{S_{\text{lo}}^{G1}} = \sigma_{S_G} + \sigma_{A_G^1} - \frac{1}{\theta}\left(\ln\left(\frac{\epsilon'}{2}\right) + \ln(1 - e^{-\theta\delta})\right). \tag{50}$$

The envelope of the leftover service of the SGL service for Poisson traffic can be written as

$$\rho_{S_{\text{lo}}^{G1}} = \rho_{S_G} - \rho_{A_G^2}, \tag{51}$$

$$\sigma_{S_{\text{lo}}^{G1}} = \sigma_{S_G} + \sigma_{A_G^2}, \tag{52}$$

TABLE II
FACILITY PARAMETERS

	Altitude (km)	Latitude (°)	Longitude (°)	Inclination (°)
DRS1	35779.36	0.89	176.75	2.5
DRS2	35773.53	-0.01	10.57	0.0
DRS3	35779.10	-1.26	76.94	2.7
USs	900	-	-	98
GS1	-	40	117	-
GS2	-	40	75	-
GS3	-	18	109	-
GS4	-	25	103	-
GS5	-	68	21	-

where $\rho_{A_G^2} = (1-\eta)/\eta\rho_{A_U^2}$, $\sigma_{A_G^2} = (1-\eta)/\eta\sigma_{A_U^2}$. Moreover, because of the intermittent characteristic of the SGL from the US to the GS and refer to Theorem 3, we have the service envelope of S_G as

$$\rho_{S_G} = \frac{1}{\theta}\ln(sp(\boldsymbol{D_s}(\theta)\boldsymbol{T_s})), \tag{53}$$

where $\boldsymbol{D_s}(\theta)$ is defined as

$$\boldsymbol{D_s}(\theta) = \text{Diag}(M_{f_{\boldsymbol{C}}}(\theta)), \tag{54}$$

where Diag$(\cdot)$ indicates the diagonal matrix, $M_{f_{\boldsymbol{C}}}(\theta)$ is the MGF of $f_{\boldsymbol{C}}$ (see Appendix B) and $\sigma_{S_G} = 0$. The derivation of the delay bound of the MMOO traffic follows the similar procedure that is omitted here to save space.

V. NUMERICAL RESULTS

A. Simulation Configuration

1) Satellite Parameters: The main software and simulation platforms used in the simulation are the STK and matlab. Herein, the configuration parameters of the scenario facilities in SDRNs including the orbital parameters of USs and DRSs, the coordinates of the GSs, and the access relationships among the facilities are all imported from the STK. On this basis, the traffic randomness, theoretical results, and simulations of the queuing process are achieved by matlab. The simulation scenario consists of eighty LEO USs, one GEO DRS network that is constituted by three DRSs to support seamless data relay, and five GSs for the data receiving. USs are distributed in eight planes with height 900 km and inclination 98°. Ten USs are uniformly deployed in each plane. Assume that the US can access any one of the DRSs in its field of vision and the DRS can access any one of the GSs in its field of vision. The detail parameters can be found in Table II.

2) Traffic Parameters: Without loss of generality, we set the arrival rate of the Poisson traffic as 20 packets per second. Each packet size is 1 Mbits. Meanwhile, the peak rate, bursty length and mean rate of the MMOO traffic are set to be 60 Mbps, 10 minutes and 20 Mbps, respectively. The distribution ratios to heterogeneous links are both 0.5, i.e., $\eta = 0.5$. For the ISL and SGL of the cascaded links, $K = [7.78, 6.99]$ dB, $\Omega = 1 + K$, $\gamma = 2$, $[P_{t_{UD}}, P_{t_{DG}}] = [10, 100]$ W, $N_0 = [-110, -100]$ dBm, bandwidth $[B_{UD}, B_{DG}] = [7.43, 32.75]$ MHz. For the discontinuous link, except $P_{t_{UG}} = 10$ W and $B_{UG} = 11.39$ MHz, the others are same as those of the SGL of the cascaded links [41], [42], [49]. The aggregated traffic from other multiplexed USs can be either the Poisson traffic, MMOO traffic or their combination. The violation probabilities ϵ' for the cascaded links and discontinuous link are set as $[0.5, 0.1]$. The initial value of the backlog is set as zero at the start time. The simulation period starts from 1 Jan. 2020 00:00:00 to 30 Jan. 2020 00:00:00.

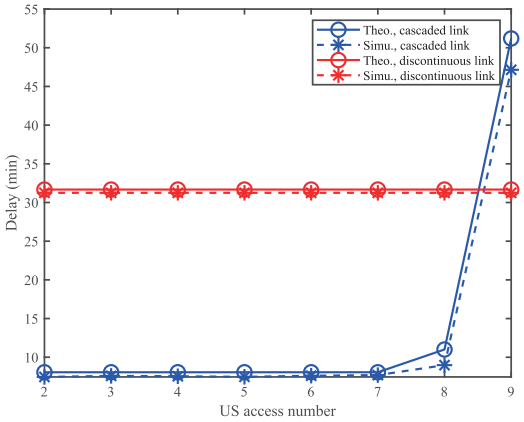

Fig. 6. The delay v.s. US access number when the multiplexing flow is the Poisson traffic.

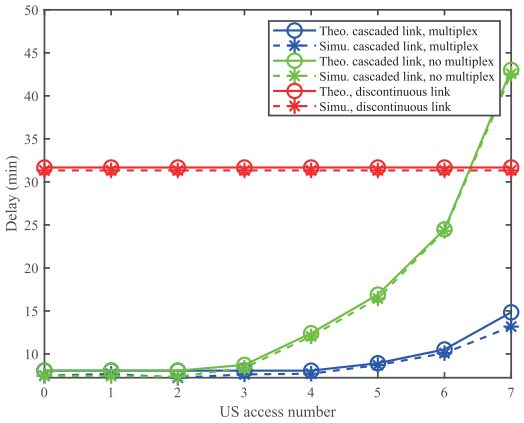

Fig. 7. The delay v.s. US access number when the multiplexing flow is the MMOO traffic.

B. Performance Evaluation

From Fig. 6, we can observe the relationship between the delay of the cascaded links and discontinuous link, and the US access number when the multiplexing flow is the Poisson traffic. Because each US has its own SGL so that the delay of the discontinuous link does not vary with the increase of the US access number. However, due to the intermittent connectivity of the discontinuous link, the backlog in the US buffer can only be transmitted during the active period of the SGL, which causes the dramatic quantity of the delay about 31 min. Besides, the theoretical results well bound the simulation ones. As for the cascaded links, we can see that the delay starts to grow until the US access number exceeds 7. Moreover, the growth rate soars when the access number is greater than 8. These phenomena reflect an important conclusions that the system begins to perform a obvious variation

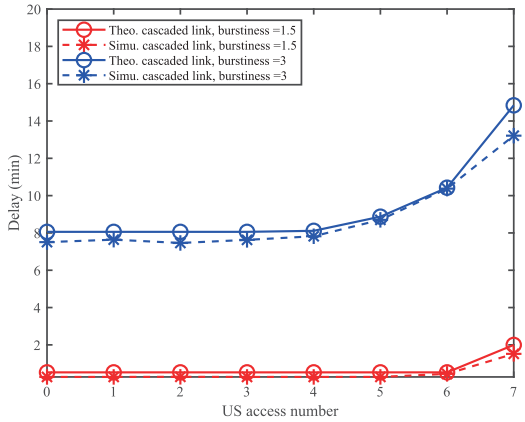

Fig. 8. The delay v.s. US access number when the multiplexing flow is the MMOO traffic.

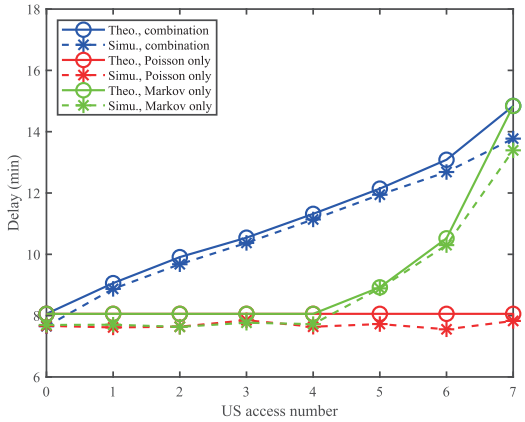

Fig. 9. The delay v.s. US access number with combined multiplexing traffic.

in delay for the non-bursty Poisson traffic only when the link occupancy approaches saturation. Then, we can find that the delay of the cascaded links does not start from zero. The underlying rationale is that the input of the US is integrated heterogeneous traffic that is constituted by the Poisson traffic and MMOO traffic. Because of the burstiness of the MMOO traffic and the existence of the Poisson traffic, the peak rate of such mixed traffic exceeds the instantaneous rate of the link service rate, which leads to the stochastic delay. Compared with the results in Fig. 8 where the burstiness of the MMOO traffic is low, the related delay begins from zero, which reveals the impact of the traffic burstiness. Moreover, the relatively high link occupancy of the ISL between the US and DRS is another essential factor. Comparing the theoretical results to the simulation results, the theoretical bound is effective.

In Fig. 7, we can see the relationship between the delay of the cascaded links and discontinuous link, and the US access number when the multiplexing flow is the MMOO traffic. Similarly, since each US has its own SGL so that the delay of the discontinuous link does not vary with the increase of the US access number. Due to the intermittent connectivity of the SGL, the data offloading only happens in the active period so that the related delay is relatively high. As for the cascaded links with the traffic multiplexing, it can be seen that the multiplexing performance of the MMOO traffic is totally different from that of the Poisson traffic. Intuitively, the mean delay of the cascaded links presents an uptrend with the increase of the US access number. Moreover, the slopes of the delay curves all show an exponential rise. This phenomenon points out that the effect of the traffic burstiness on the system delay is obvious, and exhibits an exponential feature with the increase of the US access number. When the number of accessed USs reaches 7, the related delay is about 13 min. When the link occupancy becomes nearly saturated, the SNC based bound appears a small gap. The underlying rationale is that the delay bound becomes wider since the intersection time of arrival and service envelopes becomes longer when link occupancy approaches saturation. Furthermore, we simulate the same number of accessed USs with same traffic but no multiplexing. It can be seen that the related delay is much larger, which implies the disappearance of the multiplexing gain.

As shown in Fig. 8, we study the relationship between the delay and US access number when the multiplexing flow is the MMOO type with different burstiness. Herein, we fix the mean rate of each MMOO traffic. According to the definition of the traffic burstiness that is the ratio of the peak rate to the mean rate, we turn the peak rate and the burst length into 15 Mbps and 20 min, respectively. Correspondingly, the burstiness becomes 1.5. From the figure, we can observe that the mean delay of the cascaded links does not make any difference until the US access number approaches 7. The related delay varies from near zero to about 1.8 min. This is because the occupancy of the primary ISL from the US to DRS becomes lower compared with that when the traffic burstiness is 3. As a result, fewer backlogs queue in the buffer during the on state of the MMOO traffic. Meanwhile, the overall peak rate of the multiplexing traffic flow is smaller than the link capacity of the secondary ISL from the DRS to the GS. Because of above two main reasons, not only the absolute value of the delay is relatively smaller but also its increasing rate is relatively slower than those when the burstiness is 3.

Finally, we investigate the effect of the heterogeneous traffic combination on the delay performance of the cascaded links. We set the overall US access number as 7 and increase the number of USs carrying the MMOO traffic from 0 to 7. Correspondingly, the number of USs carrying the Poisson traffic decreases from 7 to 0. Then, we conduct the simulation and compare the simulation results to the ones with USs only carrying Poisson traffic and USs only carrying MMOO traffic. From the theoretical results and simulation results shown in Fig. 9, it can be seen that the delay of the cascaded links with the combination traffic performs a nearly linear uptrend, which is totally different from the trend variations of the delay with only Poisson traffic or with only MMOO traffic. The underlying rationale is that the combination of such two kinds of traffic will enhance the burstiness of the single multiplexing MMOO traffic although the Poisson traffic does not have the burstiness itself. Moreover, the uptrend implies that the effect of the MMOO traffic on the delay is greater than that of the Poisson traffic. The small divergence of the theoretical results and simulation results when the aggregated traffic number is 7 indicates the delay bound becomes wider

since the intersection time of arrival and service envelopes becomes longer when link occupancy approaches saturation.

C. Result Implementation

Our analytical method can provide the clear guidance to the link selection and access capacity of cascaded satellites under specific stochastic delay requirements, especially under the heterogeneous traffic and transmission links. From Fig. 6 and Fig. 7, we are intuitively aware of the delay variation influenced by multiplexing a certain kind of traffic. Moreover, from Fig. 8, we can figure out the effects of the traffic burstiness on system delay. Furthermore, from Fig. 9, we can get the knowledge of the feasible selection of the traffic type and the related US number that can be accessed. For example, if the delay of the cascaded links is required to be less than 11 min, the feasible combinations of the heterogeneous traffic type and number are $(0, [1, \cdots, 7])$, $(1, 6)$, $(2, 5)$, $(3, 4)$, and $([0, \cdots, 6], 0)$ where the first value and the second value represent the aggregated number of the MMOO traffic and the Poisson traffic, respectively. Furthermore, the traffic type is not limited to the Poisson type or the MMOO type but oriented to any type whose MGF exists, which reflects the universality of our method in the aspect of the stochastic delay analysis.

VI. Conclusion

In this paper, we focused on the stochastic delay analysis to satisfy the delay requirement of the space traffic and guide the network configuration in SDRNs. Nevertheless, the complex data offloading process and the heterogeneity of the traffic and links brought great challenges to the stochastic delay analysis. To accurately model the data offloading process in SDRNs, we established a series-parallel queuing model with through and cross traffic while considering the propagation delay. Based on this queuing model, we respectively proposed a propagation delay embedded min-plus convolution method based on stochastic network calculus and a Markov chain method based on Monte Carlo to depict the introduced leftover services of the heterogeneous links received by the per-flow traffic in an aggregate. To eliminate the impacts of the heterogeneity, we uniformly characterized the arrivals and leftover services by their MGFs, and shielded the heterogeneity by deriving the envelopes of the arrivals and leftover services with the help of MGFs, Chernoff bound and union bound. According to the geometric relationship between the envelopes of the arrivals and leftover services, we derived the upper bounds of the stochastic delay, which provided the guidance to the network configuration in the light of the delay requirement of the space traffic. Finally, we conducted numerous simulations to verify the effectiveness of the theoretical analysis and offered some use cases to illustrate the practical applications of our methods.

For the future work, we plan to undertake the system modeling and performance analysis in ultra-dense satellite networks. There are large amounts of inter-satellite links among adjacent satellites so that the network topology will become more complex. As a result, the traffic flows will face the problems of the routing and multiple hops, which raises much more difficulties to the system modeling and performance analysis. Moreover, combining with the heterogeneous traffic and links, the challenges may be further amplified. However, we believe that these challenges will finally be solved properly in the near future.

Appendix A
Proof of Theorem 1

Proof: Suppose that the aggregate traffic from other multiplexed USs comes starting from d_{UD} and rewrite the both sides of (9), we have

$$\begin{aligned} &D_D^1(d_{UD},t)+D_D^2(d_{UD},t)+D_{\text{agg}}(d_{UD},t)\\ &\quad\geq \inf_{0\leq\tau_2\leq t-d_{UD}}\{A_D^1(\tau_2+d_{UD})+A_D^2(\tau_2+d_{UD})\\ &\qquad+A_{\text{agg}}(\tau_2+d_{UD})+S_D(\tau_2+d_{UD},t)\}. \end{aligned} \tag{55}$$

For the ease of description, we define $D_D^{12}(d_{UD},t) \triangleq D_D^1(d_{UD},t)+D_D^2(d_{UD},t)$ and $A_D^{12}(\tau_2+d_{UD}) \triangleq A_D^1(\tau_2+d_{UD})+A_D^2(\tau_2+d_{UD})$. Moving $D_{\text{agg}}(d_{UD},t)$ to the right-hand side of the inequation, we can further have

$$\begin{aligned} &D_D^{12}(d_{UD},t)\\ &\quad\geq \inf_{0\leq\tau_2\leq t-d_{UD}}\{A_D^{12}(\tau_2+d_{UD})+A_{\text{agg}}(\tau_2+d_{UD})\\ &\qquad+S_D(\tau_2+d_{UD},t)\}-D_{\text{agg}}(d_{UD},t)\\ &\quad= \inf_{0\leq\tau_2\leq t-d_{UD}}\{A_D^{12}(\tau_2+d_{UD})+S_D(\tau_2+d_{UD},t)\\ &\qquad-(D_{\text{agg}}(d_{UD},t)-A_{\text{agg}}(\tau_2+d_{UD}))\}. \end{aligned} \tag{56}$$

Because of $A_{\text{agg}}(t)\geq D_{\text{agg}}(d_{UD},t)$, (56) can be further written as

$$\begin{aligned} &D_D^{12}(d_{UD},t)\\ &\quad\geq \inf_{0\leq\tau_2\leq t-d_{UD}}\{A_D^{12}(\tau_2+d_{UD})+[S_D(\tau_2+d_{UD},t)\\ &\qquad-(A_{\text{agg}}(t)-A_{\text{agg}}(\tau_2+d_{UD}))]_+\}\\ &\quad= \inf_{0\leq\tau_2\leq t-d_{UD}}\{A_D^{12}(\tau_2+d_{UD})+[S_D(\tau_2+d_{UD},t)\\ &\qquad-A_{\text{agg}}(\tau_2+d_{UD},t)]_+\}, \end{aligned} \tag{57}$$

where $[\cdot]_+=\max(\cdot,0)$. Intuitively, we define the concept of the leftover service $S_{\text{lo}}^{D12}(t)$ of the secondary service for Poisson traffic and MMOO traffic that means the transmission service separated from the overall service for the specific traffic arrival process among the multiplexed traffic flows. Hence, the expression of the leftover service can be written as

$$S_{\text{lo}}^{D12}(t)=[S_D(t)-A_{\text{agg}}(t)]_+. \tag{58}$$

With the knowledge of $A_D^{12}(t+d_{UD})=D_U(t)$, and further substitute (8) into (57), we have, $\exists 0\leq\tau_1\leq\tau_2\leq t$,

$$\begin{aligned} D_D^{12}(t)&=D_D^{12}(d_{UD},t)\\ &\geq \inf_{0\leq\tau_2\leq t-d_{UD}}\{\inf_{0\leq\tau_1\leq\tau_2}\{A_U(\tau_1)+S_U(\tau_1,\tau_2)\}\\ &\quad+[S_D(\tau_2+d_{UD},t)-A_{\text{agg}}(\tau_2+d_{UD},t)]_+\}. \end{aligned} \tag{59}$$

Furthermore, we have

$$\begin{aligned} D_D^{12}(t)&=D_D^{12}(d_{UD},t)\\ &\geq \inf_{0\leq\tau_2\leq t-d_{UD}}\{\inf_{0\leq\tau_1\leq\tau_2}\{A_U(\tau_1)+S_U(\tau_1,\tau_2)\}\\ &\quad+S_{\text{lo}}^{D12}(\tau_2+d_{UD},t)\}. \end{aligned} \tag{60}$$

Note that cumulative process $S_U(t)$ is stationary because s_τ^U is independent identically distributed in each τ, the above formula can be further derived as

$$\begin{aligned} &D_D^{12}(t) \\ &\quad\geq \inf_{0\leq\tau_2\leq t-d_{UD}}\{\inf_{0\leq\tau_1\leq\tau_2}\{A_U(\tau_1) \\ &\qquad + S_U(\tau_1+d_{UD},\tau_2+d_{UD})\}+S_{\text{lo}}^{D12}(\tau_2+d_{UD},t)\} \\ &\quad= \inf_{0\leq\tau_1\leq t-d_{UD}}\{A_U(\tau_1) \\ &\qquad + \inf_{\tau_1\leq\tau_2\leq t-d_{UD}}\{S_U(\tau_1+d_{UD},\tau_2+d_{UD}) \\ &\qquad + S_{\text{lo}}^{D12}(\tau_2+d_{UD},t)\}\} \\ &\quad= \inf_{0\leq\tau_1\leq t-d_{UD}}\{A_U^1(\tau_1)+S_U\otimes S_{\text{lo}}^{D12}(\tau_1+d_{UD},t)\}, \end{aligned} \tag{61}$$

where $S_U\otimes S_{\text{lo}}^{D12}(\tau_1+d_{UD},t)\triangleq\inf_{\tau_1\leq\tau_2\leq t-d_{UD}}\{S_U(\tau_1+d_{UD},\tau_2+d_{UD})+S_{\text{lo}}^{D12}(\tau_2+d_{UD},t)\}\}$, called the min-plus convolution operation of S_U and S_{lo}^{D12}. Rewrite the both sides of (61), we have

$$\begin{aligned} D_D^1(t)+D_D^2(t)\geq &\inf_{0\leq\tau_1\leq t-d_{UD}}\{A_U^1(\tau_1)+A_U^2(\tau_1) \\ &+S_U\otimes S_{\text{lo}}^{D12}(\tau_1+d_{UD},t)\}. \end{aligned} \tag{62}$$

Moving $D_D^2(t)$ to the right-hand side of the inequation, we can further have

$$\begin{aligned} D_D^1(t)\geq &\inf_{0\leq\tau_1\leq t-d_{UD}}\{A_U^1(\tau_1)+A_U^2(\tau_1) \\ &+S_U\otimes S_{\text{lo}}^{D12}(\tau_1+d_{UD},t)\}-D_D^2(t) \\ =&\inf_{0\leq\tau_1\leq t-d_{UD}}\{A_U^1(\tau_1)+A_U^2(\tau_1) \\ &+S_U\otimes S_{\text{lo}}^{D12}(\tau_1+d_{UD},t)-D_D^2(t)\} \\ =&\inf_{0\leq\tau_1\leq t-d_{UD}}\{A_U^1(\tau_1)+S_U\otimes S_{\text{lo}}^{D12}(\tau_1+d_{UD},t) \\ &-(D_D^2(t)-A_U^2(\tau_1))\}. \end{aligned} \tag{63}$$

Because of $A_U^2(t-d_{UD})\geq D_U^2(t-d_{UD})=A_D^2(t)\geq D_D^2(t)$, (63) can be further written as

$$\begin{aligned} D_D^1(t)\geq &\inf_{0\leq\tau_1\leq t-d_{UD}}\{A_U^1(\tau_1)+[S_U\otimes S_{\text{lo}}^{D12}(\tau_1+d_{UD},t) \\ &-(A_U^2(t-d_{UD})-A_U^2(\tau_1))]_+\} \\ =&\inf_{0\leq\tau_1\leq t-d_{UD}}\{A_U^1(\tau_1)+[S_U\otimes S_{\text{lo}}^{D12}(\tau_1+d_{UD},t) \\ &-A_U^2(\tau_1,t-d_{UD})]_+\}. \end{aligned} \tag{64}$$

Intuitively, we define the concept of the leftover service $S_{\text{lo}}^{UD1}(d_{UD},t)$ of the serial service that means the per-flow service received by traffic $A_U^1(t)$ in the multiplexed traffic flows. Hence, the expression of the leftover service can be written as

$$S_{\text{lo}}^{UD1}(d_{UD},t)=[S_U\otimes S_{\text{lo}}^{D12}(d_{UD},t)-A_U^2(t-d_{UD})]_+. \tag{65}$$

So far, we have derived the relationship between the final departure process of Poisson traffic and its arrival process, primary leftover service process as well as secondary leftover service process in the cascaded links. As for the leftover service of MMOO traffic in an aggregate, the derivation follows the similar way. □

APPENDIX B
PROOF OF THEOREM 2

Proof: For the serial service for Poisson traffic, we define its MGF during $\tau\in[d_{UD},t]$ as

$$\begin{aligned} &M_{S_{\text{lo}}^{UD1}}(-\theta,t-\tau) \\ &\quad\triangleq\mathbb{E}\{e^{-\theta S_{\text{lo}}^{UD1}(\tau,t)}\} \\ &\quad=\mathbb{E}\{e^{-\theta[\inf_{\tau\leq v\leq t}\{S_U(\tau,v)+S_{\text{lo}}^{D12}(v,t)\}-A_U^2(t-\tau)]_+}\} \\ &\quad\leq\mathbb{E}\{e^{\theta A_U^2(t-\tau)}\}\sum_{v=\tau}^{t}\mathbb{E}\{e^{-\theta S_U(\tau,v)}\}\mathbb{E}\{e^{-\theta S_{\text{lo}}^{D12}(v,t)}\} \\ &\quad=M_{A_U^2}(\theta,t-\tau)\sum_{v=0}^{t-\tau}M_{S_U}(-\theta,v)M_{S_{\text{lo}}^{D12}}(-\theta,t-\tau-v) \\ &\quad=M_{A_U^2}(\theta,t-\tau)M_{S_U}*M_{S_{\text{lo}}^{D12}}(-\theta,t-\tau), \end{aligned} \tag{66}$$

where $*$ is algebra convolution operation. The right-hand side of (66) has $(t-\tau+1)$ items. Moreover, we define the MGF envelope as

$$\begin{aligned} &M_{S_{\text{lo}}^{UD1}}(-\theta,t-\tau) \\ &\quad\leq e^{\theta(\rho_{A_U^2}(t-\tau)+\sigma_{A_U^2})}(t-\tau+1) \\ &\qquad\cdot e^{-\theta(\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})(t-\tau)-\sum_{i=1}^{2}\sigma_{S_i})} \\ &\quad= e^{-\theta((\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)-(\sigma_{S_U}+\sigma_{S_{\text{lo}}^{D12}}+\sigma_{A_U^2}))} \\ &\qquad\cdot(t-\tau+1), \end{aligned} \tag{67}$$

where $\rho_{S_U}>0$, $\rho_{S_{\text{lo}}^{D12}}>0$ and $\sigma_S\geq 0$. Then, for $\forall\tau\in[0,t]$, we have the related EBF as

$$\begin{aligned} P\{S_{\text{lo}}^{UD1}(\tau,t)<(\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)-b_{S_{\text{lo}}^{UD1}}\} \\ \leq\epsilon_{S_{\text{lo}}^{UD1}}(b_{S_{\text{lo}}^{UD1}}). \end{aligned} \tag{68}$$

By introducing a violation probability, the deterministic envelope is relaxed to the stochastic version that makes the envelope tighter. The expression of the violation probability is defined as

$$\epsilon_{S_{\text{lo}}^{UD1}}(b_{S_{\text{lo}}^{UD1}})=\alpha e^{-\theta b_{S_{\text{lo}}^{UD1}}}, \tag{69}$$

where $b\geq 0$ and $\alpha\geq 0$. With the help of the Chernoff lower bound that is expressed as

$$P\{X\leq x\}\leq e^{\theta x}\mathbb{E}\{e^{-\theta X}\}, \tag{70}$$

(68) can be further derived as

$$\begin{aligned} &P\{S_{\text{lo}}^{UD1}(\tau,t)<(\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)-b_{S_{\text{lo}}^{UD1}}\} \\ &\quad\leq e^{\theta((\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)-b_{S_{\text{lo}}^{UD1}})}\mathbb{E}\{e^{-\theta S_{\text{lo}}^{UD1}(\tau,t)}\}. \end{aligned} \tag{71}$$

Combining (71) and (67), we have

$$\begin{aligned} &P\{S_{\text{lo}}^{UD1}(\tau,t)<(\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)-b_{S_{\text{lo}}^{UD1}}\} \\ &\leq(t-\tau+1)e^{\theta((\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)-b_{S_{\text{lo}}^{UD1}})} \\ &\quad\cdot e^{-\theta((\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)-(\sigma_{S_U}+\sigma_{S_{\text{lo}}^{D12}}+\sigma_{A_U^2}))} \\ &=(t-\tau+1)e^{-\theta b_{S_{\text{lo}}^{UD1}}}e^{\theta(\sigma_{S_U}+\sigma_{S_{\text{lo}}^{D12}}+\sigma_{A_U^2})}. \end{aligned} \tag{72}$$

Comparing (72) with (68), we have

$$\alpha=e^{\theta(\sigma_{S_U}+\sigma_{S_{\text{lo}}^{D12}}+\sigma_{A_U^2})}. \tag{73}$$

In order to extend the envelope in (68) corresponding to the arbitrary and fixed τ to a more general form that is suitable to the whole domain of definition of τ, we have

$$P\{\exists\tau\in[d_{UD},t]: S_{\text{lo}}^{UD1}(\tau,t)<\rho'_{S_{\text{lo}}^{UD1}}(t-\tau)-b_{S_{\text{lo}}^{UD1}}\} \le \epsilon'_{S_{\text{lo}}^{UD1}}(b_{S_{\text{lo}}^{UD1}}). \quad (74)$$

With the help of the Union lower bound that is expressed as

$$P\{\exists i: X_i\le x\}\le\sum_i P\{X_i\le x\}, \quad (75)$$

(74) can be further derived as

$$P\{\exists\tau\in[d_{UD},t]: S_{\text{lo}}^{UD1}(\tau,t)<\rho'_{S_{\text{lo}}^{UD1}}(t-\tau)-b_{S_{\text{lo}}^{UD1}}\} \le\sum_{\tau=0}^{t}P\{S_{\text{lo}}^{UD1}(\tau,t)<\rho'_{S_{\text{lo}}^{UD1}}(t-\tau)-b_{S_{\text{lo}}^{UD1}}\}. \quad (76)$$

Let $\rho'_{S_{\text{lo}}^{UD1}}(t-\tau)=(\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)-\delta(t-\tau)$ where $\delta(t-\tau)>0$ that can be optimized. Combining (72) and (76), we have

$$\begin{aligned}
&P\{\exists\tau\in[d_{UD},t]: S_{\text{lo}}^{UD1}(\tau,t)<\rho'_{S_{\text{lo}}^{UD1}}(t-\tau)-b_{S_{\text{lo}}^{UD1}}\}\\
&\le\sum_{\tau=0}^{t}P\{S_{\text{lo}}^{UD1}(\tau,t)<(\min(\rho_{S_U},\rho_{S_{\text{lo}}^{D12}})-\rho_{A_U^2})(t-\tau)\\
&\quad-\delta(t-\tau)-b_{S_{\text{lo}}^{UD1}}\}\\
&\le\sum_{\tau=0}^{t}(t-\tau+1)e^{-\theta(b_{S_{\text{lo}}^{UD1}}+\delta(t-\tau))}e^{\theta(\sigma_{S_U}+\sigma_{S_{\text{lo}}^{D12}}+\sigma_{A_U^2})}\\
&\le e^{-\theta b_{S_{\text{lo}}^{UD1}}}e^{\theta(\sigma_{S_U}+\sigma_{S_{\text{lo}}^{D12}}+\sigma_{A_U^2})}\sum_{\tau=0}^{\infty}(\tau+1)e^{-\theta\delta\tau}\\
&=\frac{e^{-\theta b_{S_{\text{lo}}^{UD1}}}e^{\theta(\sigma_{S_U}+\sigma_{S_{\text{lo}}^{D12}}+\sigma_{A_U^2})}}{(1-e^{-\theta\delta})^2}. \quad (77)
\end{aligned}$$

Recall (74), the expression of $\epsilon'(b_{S_{\text{lo}}^{UD1}})$ is

$$\epsilon'(b_{S_{\text{lo}}^{UD1}})=\frac{e^{-\theta b_{S_{\text{lo}}^{UD1}}}e^{\theta(\sigma_{S_U}+\sigma_{S_{\text{lo}}^{D12}}+\sigma_{A_U^2})}}{(1-e^{-\theta\delta})^2}. \quad (78)$$

So far, we have proved the theorem. □

APPENDIX C
PROOF OF THEOREM 3

Proof: According to (19), the definition of the MGF, and the backward equation, we can find that

$$\begin{aligned}
&\mathbb{E}\{e^{\theta A_U^2(t)}|f_{\boldsymbol{B}}(1)=i\}\\
&=\mathbb{E}\{e^{\theta A_U^2(1)}|f_{\boldsymbol{B}}(1)=i\}\mathbb{E}\{e^{\theta(A_U^2(t)-A_U^2(1))}|f_{\boldsymbol{B}}(1)=i\}\\
&=M_{f_{B_i}}(\theta)\sum_{j=1}^{2}\mathbb{E}\{e^{\theta(A_U^2(t)-A_U^2(1))}|f_{\boldsymbol{B}}(2)=j,f_{\boldsymbol{B}}(1)=i\}\\
&\quad\cdot P(f_{\boldsymbol{B}}(2)=j|f_{\boldsymbol{B}}(1)=i)\\
&=M_{f_{B_i}}(\theta)\sum_{j=1}^{2}\mathbb{E}\{e^{\theta(A_U^2(t)-A_U^2(1))}|f_{\boldsymbol{B}}(2)=j\}p_{ij}^a\\
&=M_{f_{B_i}}(\theta)\sum_{j=1}^{2}\mathbb{E}\{e^{\theta(A_U^2(t-1))}|f_{\boldsymbol{B}}(1)=j\}p_{ij}^a. \quad (79)
\end{aligned}$$

Let $\phi(\theta,t)=\left[\mathbb{E}(e^{\theta A_U^2(t)}|f_{\boldsymbol{B}}(1)=1),\mathbb{E}(e^{\theta A_U^2(t)}|f_{\boldsymbol{B}}(1)=2)\right]$ and $\phi(\theta,t)^T=\boldsymbol{D_a}(\theta)\boldsymbol{T_a}\phi(\theta,t-1)^T$ with the initial condition

$$\phi(\theta,1)^T=\boldsymbol{D_a}(\theta)\mathbf{1}_2, \quad (80)$$

where $\mathbf{1}_2$ is a two-dimensional column vector with all 1 elements. Furthermore, we have

$$M_{A_U^2}(\theta,t)=\boldsymbol{\pi}\phi(\theta,t)^T=\boldsymbol{\pi}(\boldsymbol{D_a}(\theta)\boldsymbol{T_a})^{t-1}\boldsymbol{D_a}(\theta)\mathbf{1}_2. \quad (81)$$

Since $sp(\boldsymbol{D_a}(\theta)\boldsymbol{T_a})$ is the spectral radius of the matrix $\boldsymbol{D_a}(\theta)\boldsymbol{T_a}$, for every $\xi>0$ there exists a constant $\sigma_\xi(\theta)<\infty$ so that every element of the matrix $(\boldsymbol{D_a}(\theta)\boldsymbol{T_a})^t$ is bounded by $\sigma_\xi(\theta)(sp(\boldsymbol{D_a}(\theta)\boldsymbol{T_a})+\xi)^t$ [50]. Recall (81), we can intuitively have

$$\lim_{t\to\infty}\sup\frac{1}{\theta t}\ln\mathbb{E}\{e^{A_U^2(t)}\}\le\frac{1}{\theta}\ln(sp(\boldsymbol{D_a}(\theta)\boldsymbol{T_a})+\xi). \quad (82)$$

As ξ is arbitrary, let $\xi\to0$, we have $\sigma_{A_U^2}=0$ and

$$\rho_{A_U^2}=\frac{1}{\theta}\ln sp(\boldsymbol{D_a}(\theta)\boldsymbol{T_a}). \quad (83)$$

So far, we have proved the theorem. □

REFERENCES

[1] M. Sheng, Y. Wang, J. Li, R. Liu, D. Zhou, and L. He, "Toward a flexible and reconfigurable broadband satellite network: Resource management architecture and strategies," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 127–133, Aug. 2017.

[2] M. Sheng, D. Zhou, R. Liu, Y. Wang, and J. Li, "Resource mobility in space information networks: Opportunities, challenges, and approaches," *IEEE Netw.*, vol. 33, no. 1, pp. 128–135, Jan. 2019.

[3] Z. Zhang *et al.*, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[4] K. David and H. Berndt, "6G vision and requirements: Is there any need for beyond 5G?" *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 72–80, Sep. 2018.

[5] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-Air-Ground integrated network: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2714–2741, May 2018.

[6] M. Harris, "Tech giants race to build orbital Internet [News]," *IEEE Spectr.*, vol. 55, no. 6, pp. 10–11, Jun. 2018.

[7] J. Foust, "SpaceX's space-Internet woes: Despite technical glitches, the company plans to launch the first of nearly 12,000 satellites in 2019," *IEEE Spectr.*, vol. 56, no. 1, pp. 50–51, Jan. 2019.

[8] J. Du, C. Jiang, Q. Guo, M. Guizani, and Y. Ren, "Cooperative Earth observation through complex space information networks," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 136–144, Apr. 2016.

[9] H. K. Ramapriyan, "The role and evolution of NASA's earth science data systems," NASA Goddard Space Flight Center, Greenbelt, MD, USA, Tech. Rep. GSFC-E-DAA-TN24713, 2015.

[10] L. Wang, C. Jiang, L. Kuang, S. Wu, and S. Guo, "TDRSS scheduling algorithm for non-uniform time-space distributed missions," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–6.

[11] J. J. Gramling and N. G. Chrissotimos, "Three generations of NASA's tracking and data relay satellite system," in *Proc. AIAA*, ESA, Heidelberg, Dec. 2008, pp. 1–11.

[12] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. London, U.K.: Springer, 2009.

[13] M. Fidler, "Survey of deterministic and stochastic service curve models in the network calculus," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 1, pp. 59–86, 1st Quart., 2010.

[14] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–105, 1st Quart., 2015.

[15] L. Lei *et al.*, "Stochastic delay analysis for train control services in next-generation high-speed railway communications system," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 48–64, Jan. 2016.

[16] R. Liu, M. Sheng, K.-S. Lui, X. Wang, D. Zhou, and Y. Wang, "Capacity analysis of two-layered LEO/MEO satellite networks," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, Glasgow, U.K., May 2015, pp. 1–5.
[17] H. Nishiyama, Y. Tada, N. Kato, N. Yoshimura, M. Toyoshima, and N. Kadowaki, "Toward optimized traffic distribution for efficient network capacity utilization in two-layered satellite networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 3, pp. 1303–1313, Mar. 2013.
[18] Y. Kawamoto, H. Nishiyama, N. Kato, and N. Kadowaki, "A traffic distribution technique to minimize packet delivery delay in multilayered satellite networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3315–3324, Sep. 2013.
[19] Y. Wang *et al.*, "Multi-resource coordinate scheduling for Earth observation in space information networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 2, pp. 268–279, Feb. 2018.
[20] D. Zhou, M. Sheng, R. Liu, Y. Wang, and J. Li, "Channel-aware mission scheduling in broadband data relay satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 1052–1064, May 2018.
[21] R. Liu, M. Sheng, C. Xu, J. Li, X. Wang, and D. Zhou, "Antenna slewing time aware mission scheduling in space networks," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 516–519, Mar. 2017.
[22] T. Zhang, H. Li, S. Zhang, J. Li, and H. Shen, "STAG-based QoS support routing strategy for multiple missions over the satellite networks," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6912–6924, Oct. 2019.
[23] T. Zhang, H. Li, J. Li, S. Zhang, and H. Shen, "A dynamic combined flow algorithm for the two-commodity max-flow problem over delay-tolerant networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7879–7893, Dec. 2018.
[24] Y. Mai and P. Palmer, "Satellite imaging service analysis using queueing theory," in *Proc. AIAA/USU*, Logan, UT, USA, Aug. 2001, pp. 1–12.
[25] W. Chen, "Performance modelling of imaging service from earth observation satellites," Ph.D. dissertation, Dept. Electron. Phys. Sci., Surrey Space Centre, Univ. Surrey, Guildford, U.K., 2007.
[26] W. Chen, S. Mackin, and P. Palmer, "Performance modelling of imaging service of earth observation satellites with two-dimensional Markov chain," in *Proc. AIAA/USU*, Logan, UT, USA, Aug. 2006, pp. 1–10.
[27] W. Chen, P. Palmer, S. Mackin, and G. Crowley, "Queuing theory application in imaging service analysis for small Earth observation satellites," *Acta Astronautica*, vol. 62, nos. 10–11, pp. 623–631, May 2008.
[28] Y. Zhu, M. Sheng, J. Li, D. Zhou, and Z. Han, "Modeling and performance analysis for satellite data relay networks using two-dimensional Markov-modulated process," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3894–3907, Jun. 2020.
[29] Y. Zhu, M. Sheng, J. Li, and R. Liu, "Performance analysis of intermittent satellite links with time-limited queuing model," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2282–2285, Nov. 2018.
[30] Y. Zhu, M. Sheng, J. Li, R. Liu, Y. Wang, and K. Chi, "Traffic modeling and performance analysis for remote sensing satellite networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.
[31] Y. Hu, H. Li, Z. Chang, and Z. Han, "End-to-End backlog and delay bound analysis for multi-hop vehicular ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6808–6821, Oct. 2017.
[32] T. Liu, L. Sun, R. Chen, F. Shu, X. Zhou, and Z. Han, "Martingale theory-based optimal task allocation in heterogeneous vehicular networks," *IEEE Access*, vol. 7, pp. 122354–122366, 2019.
[33] G. Wu, W. Pedrycz, H. Li, M. Ma, and J. Liu, "Coordinated planning of heterogeneous Earth observation resources," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 46, no. 1, pp. 109–125, Jan. 2016.
[34] Y. Hu, H. Li, Z. Chang, and Z. Han, "Scheduling strategy for multimedia heterogeneous high-speed train networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3265–3279, Apr. 2017.
[35] Y. Hu, Z. Chang, H. Li, T. Ristaniemi, and Z. Han, "Service provisioning and user association for heterogeneous wireless railway networks," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3066–3078, Jul. 2017.
[36] Y. Shi, Z. Cheng, Q. You, and M. Liu, *Research on Handover Strategy of Low Orbit Spacecraft Based on Multi-beam GEO Communication Satellite*. London, U.K.: Springer, 2019.
[37] Z. Deng, B. Long, W. Lin, and J. Wang, "GEO satellite communications system soft handover algorithm based on residence time," in *Proc. 3rd Int. Conf. Comput. Sci. Netw. Technol.*, Dalian, China, Oct. 2013, pp. 834–838.
[38] H. Song, S. Liu, X. Hu, X. Li, and W. Wang, "Load balancing and QoS supporting access and handover decision algorithm for GEO/LEO heterogeneous satellite networks," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Dec. 2018, pp. 640–645.
[39] D. Bertsekas and R. Gallager, *Data Network*. Upper Saddle River, NJ, USA: Prentice-Hall, 1992.
[40] H.-W. Ferng, "Modeling of split traffic under probabilistic routing," *IEEE Commun. Lett.*, vol. 8, no. 7, pp. 470–472, Jul. 2004.
[41] J. Du, C. Jiang, J. Wang, Y. Ren, S. Yu, and Z. Han, "Resource allocation in space multiaccess systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 2, pp. 598–618, Apr. 2017.
[42] B. Di, H. Zhang, L. Song, Y. Li, and G. Y. Li, "Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 47–62, Jan. 2019.
[43] Y. Ruan, Y. Li, C.-X. Wang, R. Zhang, and H. Zhang, "Energy efficient power allocation for delay constrained cognitive satellite terrestrial networks under interference constraints," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4957–4969, Oct. 2019.
[44] A. Abdi, C. Tepedelenlioglu, M. Kaveh, and G. Giannakis, "On the estimation of the k parameter for the Rice distribution," *IEEE Commun. Lett.*, vol. 5, no. 3, pp. 92–94, Mar. 2001.
[45] A. Goldsmith, *Wireless Communications*. New York, NY, USA: Cambridge Univ. Press, 2005.
[46] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Milton Park, U.K.: Taylor & Francis, 1995.
[47] M. Wang, J. Li, Y. Jiang, and X. Di, "Stochastic performance analysis for LEO inter-satellite link based on finite-state Markov chain modeling," in *Proc. 4th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT)*, Harbin, China, Dec. 2015, pp. 1230–1235.
[48] C.-S. Chang, *Performance Guarantees in Communication Networks*. London, U.K.: Springer, 2000.
[49] M. Gerard and B. Michel, *Satellite Communications Systems*. Hoboken, NJ, USA: Wiley, 2002.
[50] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

Yan Zhu received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2015, where he is currently pursuing the Ph.D. degree in communication and information systems. He is also a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA. His research interests focus on network modeling, performance analysis, and resource allocation in wireless communication networks.

Di Zhou (Member, IEEE) received the B.E. degree in information engineering and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2013 and 2019, respectively. She was also a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Houston, from 2017 to 2018. Since 2019, she has been with the Broadband Wireless Communications Laboratory, School of Telecommunications Engineering, Xidian University, where she currently holds a faculty post-doctoral position. Her research interests include routing, resource allocation, and mission planning in space information networks.

Min Sheng (Senior Member, IEEE) received the degree from Xidian University, in 2000. She is currently a Full Professor and the Director of the State Key Laboratory of Integrated Services Networks, Xidian University. She has published over 200 refereed articles in international leading journals and key conferences in the area of wireless communications and networking. Her current research interests include space-terrestrial integration networks, intelligent wireless networks, and mobile ad hoc networks. She received China National Funds for Distinguished Young Scientists in 2018. She is an Editor for the IEEE COMMUNICATIONS LETTERS and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the Vice Chair of the IEEE Xi'an Section.

Jiandong Li (Senior Member, IEEE) received the bachelor's, master's, and Ph.D. degrees from Xidian University, in 1982, 1985, and 1991, respectively.

He has been with Xidian University since 1985. He was an Associate Professor from 1990 to 1994. He has been a Professor of Xidian University since 1994, the distinguished professor of the Chang Jiang Scholar Program, and was a Visiting Professor of Cornell University from 2002 to 2003. He was the Dean of the School of Telecommunication Engineering, Xidian University, from 1997 to 2006, the Executive Vice Dean of the Graduate School of Xidian University from 2007 to 2012, and the Vice president of Xidian University from 2012 to 2019. He was awarded the National Science Fund for Distinguished Young Scholars. He is the fellow of the China Institute of Electronics (CIE) and China Institute of Communication (CIC). He was a member of the PCN (Personal Communications Networks) specialist Group for China 863 Communication High Technology Program from 1993 to 1994 and from 1999 to 2000. He is also a member of the specialist group of the new generation of broadband wireless mobile communication networks for The Ministry of Industry and Information Technology, and the Chair of the Broadband Wireless IP Standard Work Group, China. His main research interests include mobile wireless communications, cognitive and self-organizing networks, and wireless ad-hoc networks.

Zhu Han (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, MA, USA, in 1999 and 2003, respectively.

From 2000 to 2002, he was a Research and Development Engineer of JDSU, Germantown, MA, USA. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an Assistant Professor at Boise State University, Boise, ID, USA. He is currently a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department as well as with the Computer Science Department, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, the IEEE Leonard G. Abraham Prize in the field of Communications Systems (Best Paper Award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018, AAAS fellow since 2019, and an ACM distinguished member since 2019. He has been a highly cited researcher (1%) since 2017, according to the Web of Science. He is also the winner of the 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."