SMARTSAT
COOPERATIVE RESEARCH CENTRE

TECHNICAL REPORT 3

Advanced Satellite Communications for High Rate and Dynamic Service Delivery

(SCOPING STUDY)

TECHNICAL REPORT 3

Advanced Satellite Communications for High Rate and Dynamic Service Delivery

(SCOPING STUDY)

NOVEMBER 2023

SMARTSAT COOPERATIVE RESEARCH CENTRE

MACQUARIE University

UNSW SYDNEY

THE UNIVERSITY OF SYDNEY

This report should be cited as:
SmartSat 2021, Advanced Satellite Communications for High Rate and Dynamic Service Delivery, SmartSat Technical Report 3, SmartSat, Adelaide, Australia.

Disclaimer:
This publication is provided for the purpose of disseminating information relating to scientific and technical matters. Participating organisations of SmartSat do not accept liability for any loss and/or damage, including financial loss, resulting from the reliance upon any information, advice or recommendations contained in this publication. The contents of this publication should not necessarily be taken to represent the views of the participating organisations.

Acknowledgement:
SmartSat acknowledges the contribution made by the partner groups on this SmartSat CRC project from Macquarie University, the University of New South Wales, and the University of Sydney

Contents

Executive Summary

This report was compiled for the SmartSat CRC as an output of the Scoping Study: Advanced Satellite Communications for High Rate and Dynamic Service Delivery. It is delivered in two parts. Part I is a literature survey that reviews and summarises the current state of the art in advanced satellite communications for high rate and dynamic service delivery. Part I explores opportunities in the development of future high through- put systems satellite technologies and identifies potential research and development areas for the SmartSat CRC. Part II presents research findings from the project to shed more light on the potential opportunites identified in Part I.

The Scoping Study had Six Tasks and associated deliverables. Part I of this report is the deliverable associated with Task 1. Sections 9-13 in Part II of this report are the deliverables for Tasks 2-6.

The key findings from Part I are as follows:

- The current state-of-the-art in Spot Beams is still focused on multi-horn reflector antennas which create fixed spot beams on the ground. Multibeam precoding has been considered in the literature but has not been taken to practical implementation. It is recognized that multibeam precoding is essential to achieve universal frequency reuse across all beams.
- Recent work has shown the benefits of combining multibeam precoding with user scheduling. Gains of 100% can be achieved with MMSE beamforming and geographic scheduling.
- Direct radiating active phased arrays (APAs) are highly beneficial in terms of flexi- bility and efficiency in HTS systems. However, for narrow beams in the order of few kilometers (e.g., a small mining city), the beam width has to be a small fraction of a degree at the GEO orbit. This requires APAs with very large number of antennas (in the order of 1000s). Due to the smaller wavelength of Ka band, this is possible to achieve with planar arrays which span several meters. However, wider implica- tions of the deployment of such large Ka-band APA arrays at the satellites in terms of heat generation, RF interference, weight, cost, life time of the communication payload and so on, need to be investigated.
- Serving a large number of customers in a large moving vehicle (e.g., a cruise ship, a train, an aircarft) efficiently is an important application of APAs. In such applica- tions, the beam generated by the APA has to follow the vehicle. Such a system has not been studied or implemented yet.
- Matching traffic to beams has so far only been addressed on a per cell aggregate level using mean rates of users. With APAs, adapting to traffic on hour by hour or even minute by minute basis is achievable but mechanisms to coordinate user information and resource allocation are yet to be identified. To meet 5G-type QoS requirements will require taking account of user traffic variation, not just mean rates.
- For beam assignment to gateways the current state-of-the-art is for fixed beam systems. However, with adaptive beams, association of gateway to beam is much more complex due to the increased degrees of freedom in dynamic beamforming from the satellite, including the required channel state information, and the bandwidth constraints of the feeder links.

- High throughput LEO satellite communication systems are still in their infancy. At the physical layer there are significant Doppler shifts and new delay-Doppler communication techniques are highly promising. In LEO satellite networking there is currently no systematic approach to optimize the distribution of network algo- rithms between ground and space. Similar conclusions apply to edge computing and computation offloading.

- Phased arrays in LEOs will allow much smaller beams but require a joint approach to beam steering/beam hopping/handover between LEOs, yet to be developed.

- Nonlinearity is a crucial factor in satellite communication systems. Existing compensation methods are based on single carrier systems. They are not adequate for multi-carrier OFDM based systems. Machine learning techniques are promising, but have yet to be applied to satellite nonlinear compensation.

- The impact of the shortage and instability of power supplied by arrays of solar cells on edge computing has not been considered in the literature. Power generation prediction of renewable energy sources for the edge data center has also not been considered.

- Prediction and communications for LEOs should be co-designed to optimize routing, scheduling, precoding and resource allocation using out-of-date information (due to long propation delays) and Doppler shifts. Deep learning techniques are promising but yet to be applied.

- There is no existing software-defined radio access network (RAN) architecture for 5G NTN in the literature.

Research was undertaken in Part II to address some of the gaps identified in Part I. Th key findings from Part II are as follows:

- Comparing conventional reflecting antennas with active phased arrays, the overall capacity depends on the shapes of the feedhorns in the former case, and the number of antennas in the latter. Our results show that overall capacity is comparable for similar sized spot beams, but phased arrays can be reconfigured to provide capacity to where it is needed. Both approaches benefit significantly from a digital beamforming capability, allowing full frequency re-use instead of the conventional frequency re-use factor of 1⁄4.

- Results obtained in Section 9 provide two novel approaches to dynamic spot beams. The first uses beam hopping to ensure that each beam spends more time in traffic hotspots, hence balancing per-user data rates. The second uses dynamic RF chain allocation which allocates more beams to hotspot areas all of the time. The two approaches yield similar performance results, with slightly better performance from the RF chain allocation scheme, particularly for users in lighter traffic areas.

- Using reconfigurable APAs increases rates in congested squares by a factor of about 4 as compared with fixed reflecting antennas unmatched to the traffic distribution. Digital precoding with both fixed and reconfigurable systems can increase capacity by a factor of about 2 using full frequency re-use, as compared to the corresponding system with analog spot beams and frequency re-use of 1/4.

- Both proposed dynamic spot beam schemes are heuristic, and much more research is required to understand the performance limits, and to obtain algorithms with the best possible performance, even for the simple traffic model considered in this report.

- The traffic model considered in Section 9 is very simple, with the whole geographic region divided into equal sized square regions with three different colours labelling three possible traffic levels in each square. Squares of the same colour have the same traffic level. The model is spatially inhomogeneous but time invariant. Future re- search is required to obtain more realistic, time-varying traffic models based on user activity, and where users can be mobile eg. trucks, trains, ships or planes. Questions such as how the traffic information is obtained, and how spot beams can be adapted on an appropriate time-scale to meet more detailed user-level performance targets remain to be investigated.

- Balancing data requirements for each beam by user scheduling is also considered in Section 10 where the focus is on multigroup multicast with universal frequency reuse. We propose a joint precoding and user scheduling approach based on traffic demands.

- The joint precoding and user scheduling traffic matching problem is non-convex and combinatorial. In Section 10.2, we formulate a new joint precoding and user scheduling traffic matching problem by relaxing some of the constraints, and propose solving it via an iterative approach.

- In Section 10.1, we propose a precoding algorithm that guarantees fair resource allocation among users and which is robust to phase uncertainties of channel state information. In proposed future work, a user scheduling method will be jointly studied based on the proposed precoding algorithm.

- In Section 10.5, we introduce fundamentals of OTFS, and explain its advantages over the existing modulation schemes in terms of Doppler-resilience with theoretical analysis and simulation results. It is shown that OTFS has great potential for LEO satellites with frequent and high-speed movements, and its low PAPR and low complexity is also desirable for SatCom systems.

- In Section 11, we focus on the problem of compensating for non-linearity in the high power amplifier (HPA), identified in Part I, and propose a Neural Network-based Digital Pre-Distortion (DPD) scheme. We show that this scheme can largely miti- gate the nonlinearity of the HPA. It can characterize the HPA's nonlinear behavior and the memory effect. Our NN-based DPD has a better performance compared to conventional DPD.

- Memory cells such as used in recurrent neural networks (RNNs) could be used to account for memory effects in the HPA. Time varying parameters in the HPA model could also be considered in future work. It remains to validate the effect of NN-based DPD on other parts of the communication system such as orthogonal frequency division multiplexing (OFDM), and multiple-input and multiple-output (MIMO).

- In Section 12, we show that it is feasible to convert software algorithms to a hardware and software co-design system. The next steps include converting 5G-related SDR functions to FPGA-based algorithms, optimizing the algorithms by maximizing the advantages of FPGA, and verifying and debugging the system.

- In Section 13, we propose an AI-enabled SDN architecture for 5G NTNs, attempting to fill the gap identified in Part I. We have identified the possible research problems to design and develop the architecture and have made a preliminary evaluation of the proposed architecture. We show the feasibility of the transmission of the 5G waveform over satellite networks. We develop and evaluate a deep reinforcement learning algorithm for the scheduler design and find that the delayed CSI in 5G NTNs leads to low reliability. In future work, we can design the SDN architecture to enable the programmability of 5G NTNs for different SDN applications. Then, AI algorithms can be developed to optimise the end-to-end performance over the 5G NTNs based on the SDN architectures.

Part I

Literature Review of

High-Throughput and Adaptive Satellite Communications

High Throughput Satellite (HTS) communications has recently stepped into a new era, enjoying renewed attention and a rapidly growing market in global telecommunications. New technologies are driving the opportunities, including direct radiating active phased arrays, multibeam precoding, higher frequency bands, and cloud based software defined networking. These technologies enable future flexible satellite deployments with adap- tive spot beams, traffic aware resource allocation, and integration with complementary networks including 5G mobile.

This part of the report reviews and summarises the current state of the art in ad- vanced satellite communications for high rate and dynamic service delivery. It explores opportunities in the development of future high throughput systems satellite technologies and identifies potential research and development areas for the SmartSat CRC.

The organization of the report has been guided by the following aims of the overall Scoping Study:

- Establish the potential gains from adaptive spot beams for Ka band multibeam satellites and the challenges with practical implementation under constraints on computation, latency, feeder link bandwidth, feeder link availability, inter-beam interference, and variations in traffic demands.
- Establish the potential gains from combined digital precoding and user selection for user uplink and downlink for multibeam satellites
- Establish the potential for re-configurability and modularity of satellite gateways, user terminals, and payloads in terms of coverage, orbital location, power and bandwidth by developing machine learning based physical layer and MAC techniques with new signal waveforms, novel multiple access schemes and MU-MIMO satellite technologies and their SDR implementation with FPGA technologies.
- Understand the challenges and potential for the use of higher frequency bands, including Q, V, W and optical, for both feeder links and inter-satellite links
- Establish the benefits of SDN technologies for orchestration, control, resource shar- ing and network slicing in satellite networks by developing an SDN architecture, use case scenarios and requirements for practical implementation.

In this part of the report, we review the state-of-the-art for the high-throughput and adaptive satellite communications. Section 1 provides the background to the report. Sec- tion 2 focuses on multibeam precoding techniques which are particularly useful in the context of direct radiating active phased arrays (APAs), which can be used to synthe- sise dynamic spot beams. This Section shows the benefits from combining multibeam precoding with user scheduling. Section 3 focuses on the state-of-the-art in Adaptive Spot Beams, including user scheduling, beam hopping, adaptive power and bandwidth allocation, and what has been done so far on matching spot beam capacities with traffic demands. This section is focused on GEO satellites. Section 4 considers the additional challenges arising in networks of LEO Satellites, which includes combined approaches to beam steering, beam hopping, and handover between LEOs. Challenges arising from nonlinear effects of satellite amplifiers are considered in Section 5. This Section also considers the challenges of providing physical layer security. Computing techniques in satellite systems are reviewed in Section 6 which includes the challenges of resource allo- cation in LEO networks, including consideration of time varying power generation from solar panels, and computation offloading in the dynamic environment of moving LEO constellations. In Section 7, FPGA-based software-defined radio technologies for satellite transponders are reviewed. Software-defined satellite network architectures are described in Section 8, including state-of-the-art satellite constellation design, routing and cross- layer design. A number of opportunities are identified for the application of machine learning techniques in handling multi-user interference, non-linear compensation, edge computing, and software defined networking.

Key Findings and Future Research opportunities are listed and summarized in the last subsection of each Section. In these subsections, we summarize the state-of-the-art, and identify the gaps where there are future research opportunities.

1 Background

1.1 Satellite Channel Modelling

Before a discussion of precoding and adaptive spot beams, it is worthwhile to quickly review some key properties of satellite channel models.

1.1.1 Path Loss

In the absence of interference and in “clear sky” conditions the SNR can be determined via the path loss and antenna gains using the Friis formula. Thus the path loss is determined as

$$10 \log_{10} \frac{4\pi d}{\lambda}^{2}$$

with λ determined by Ka Band frequencies, say 30 GHz, 20 GHz from the satellite to/from the user terminals to give 1 - 1 1/2 cm as the wavelength. This puts the Ka band in the low part of the mm wave spectrum. The parameter×d is the satellite distance. For a GEO satellite distance of d = 37.5 10^6m this gives a loss of 230 dB. The corresponding one-way delay is 125 msec.

Additional losses come from other factors such as rain and cloud attenuation, gas and water vapour attentuation. Deep absorption at 22.3 GHz means that this frequency is not used. The most significant factor is rain attenuation which can give rise to additional loss of as much as 25 dB. Other effects include losses from infrastructure surrounding the mobile, usually modeled using lognormal shadowing, see [1].

1.1.2 Fading

In most cases the signal has a strong LOS component so that the statistics of fading are Rician, see for example [2]. This arises from scattering from buildings and other infrastructure in the vicinity of the user terminal (UT).

A two state model for LEO satellites is described in [1] where in the shadowed state the fading is Rayleigh and in the non-shadowed state it is Rician with K-factor c = 10 for both Rayleigh and Rician. Only the proportion of time the UT is in one state or the other is specified and the dynamics are not discussed. This paper also observes that fading on the downlink is the same for all the beams (they analyse a CDMA system). This implies that the impact of multiple access interference is reduced over what would be experienced for independent fading as is experienced in terrestrial mobile networks. A similar observation is made in [2] for the uplink. Thus there is a single (Rician) fading factor for each UT, the same across all beams.

1.1.3 LEO Channel Models

The main point of difference between channel models for LEO and GEO satellite systems is the extreme Doppler shifts experienced in LEO systems due to high satellite mobility [3–5]. Hence, in this part, we will mainly focus on the Doppler shift phenomenon observed in LEO satellite communications systems, which can be as large as 40 [kHz] at speeds of 5 [km/s] (i.e., see [4]), and discuss its impact on the physical layer techniques.

An accurate characterization and estimation of Doppler shifts in LEO satellite systems is an important problem to overcome, which helps alleviating performance degradation due to frequency synchronization failures between transmitters and receivers. An early study in this direction is the work by Ali *et al.* [3] that provides a novel Doppler shift char- acterization for the forward link from LEO satellites to ground terminals. In particular, by assuming circular orbits with respect to the Earth and constant angular velocity for LEO satellites with respect to the user on the Earth, they provided an analytical curve, parametrized by the maximum elevation angle, showing the time variation of the Doppler shift within the visibility time window. This characterization of Doppler shift is important for the design of phase-locked loop circuits to lock in the transmission frequency for data decoding, especially if OFDM type modulation is used for data transmissions. Further, it can be used as a basis to perform flow control in bent-pipe LEO satellite network archi- tectures [6]. It can also be used to obtain more advanced stochastic channel models such as the finite state Markov channel model for LEO satellite systems proposed in [7]. More recently, the Doppler characterization in [3] is improved by means of statistical signal processing techniques to provide a maximum a posteriori probability estimator for the Doppler shift in [4]. The new method brings about 4 [dB] performance gain (in terms of mean squared error) in estimating Doppler shifts for LEO satellite systems, especially when the ground terminals are also mobile.

1.1.4 Multi-beam Channel Models:

Consider the downlink and the received signal, $r(t)$, of a single user in a multi-user CDMA system, [1]:

$$r(t) = AR\sum_{j=1}^{J}\sum_{k=1}^{K}\beta_{j,k}x_{j,k} + N(t) \tag{1}$$

where A is the net signal amplitude, R is the fading factor, J is the number of spot beams, K is the number of users per spot beam, $\beta_{j,k}$ is the antenna radiation from user k's signal in beam j to the single user with received signal $r(t)$. This formula is for a CDMA system but can be modified for systems with reuse and time sharing of carriers, [8]. Note that in this model the interference on the downlink is not determined by which users are scheduled together, unless there are carriers/timeslots which are left idle.

For the reverse link the received signal vector at the satellite is given via

$$\mathbf{y}(t) = \mathbf{A}\mathbf{D}(t)\mathbf{x}(t) + \mathbf{z}(t) \tag{2}$$

see [2] (9). Here we consider a system with reuse and neglect adjacent channel interference. The vector $\mathbf{x}(t) \in \mathbb{C}^{BK}$ contains the user signals, where there are B beams, each focused on a particular cell, and there are K users in each cell. The vectors of aggregate received signal and noise, respectively, are denoted by $\mathbf{y}, \mathbf{z} \in \mathbb{C}^{B}$, where B is the number of beams. The matrix of channel gains $\mathbf{A} \in \mathbb{C}^{B\times BK}$ gives the gains from a user in a cell to all of the beams. For the bth row and users in cell c produced by beam c we have the entry

$$\mathbf{a}_{b,c} = \alpha_{\ell}^{\pounds}\ {}_{\pounds\in c}$$

where $\alpha_{\ell}^{\pounds}$ denotes the gain between user £ of cell c and beam b which depends on the antenna gain and path loss for user £. The matrix $\mathbf{D} \in \mathbb{C}^{BK\times BK}$ is a diagonal matrix of independent flat fading processes, with one entry per user. In this model, the fading coefficient for each beam for a given user are the same.

These equations can be modified to reflect TDMA and the use of carriers. Observe that in this case (2) implies that the users SINR or equivalent are unknown as these depend on which users are scheduled on which timeslots/carriers. The instantaneous fading cannot be known at the receiver for a GeoStationary satellite since the delay between measurement and scheduling is too long. This means that we cannot scheduling according to instantaneous rates, as would be the case in terrestrial systems.

1.2 High Throughput Satellite Communications

The broadband high throughput satellite (HTS) systems are becoming one of the most popular promising components for SatComs. HTS is also called very-high throughput satellite (VHTS), or ultra-high throughput satellite (UHTS) in some literature, and it will be hereinafter referred to as HTS. HTS usually generates multiple narrower spot- beams with smaller coverage rather than the conventional wide beams. The multibeam technology enables a larger bandwidth of HTS and can dramatically increase the capacity of SatComs. Hence, HTS has great potential in cost-effectively delivering innovative and flexible broadband services to a large subscriber base with high quality and reliability, which establishes the cornerstone of success. The potential markets for HTS are as follows:

- Compared with the terrestrial wireless communication systems, HTS predominates in the unserved or underserved areas, such as developing countries, remote areas, offshore platforms, and regions with extreme natural geographical features.

- HTS has growing demands in transportation, such as aero and maritime connectivity, broadband delivery to trains, cruise ships and passenger aircrafts, where a large number of passengers (in the order of thousands) demand high-speed data connections, but the cellular services are unavailable or quality-limited.

- HTS can be considered to provide cellular backhaul if it is cost-competitive with terrestrial-based wireless solutions [9].

Table 1: Commercial high throughput satellite systems. In the table, “EMEA” refers to Europe, Middle East and Africa.

System	Company	Orbit	Time of Launch	Band	Total Throughput	Number of Spot-Beams	Coverage
ViaSat-1	ViaSat	GEO	10/2010	Ka	140 Gbps	72	North America
ViaSat-2		GEO	06/2017	Ka	260 Gbps	120	Americas
ViaSat-3		GEO	After 2020	Ka	1 Tbps (expected)	1000	Class 1: Americas Class 2: EMEA Class 3: Asia
Jupiter-1 (EchoStar XVII)	Hughes	GEO	07/2012	Ka	100∼140 Gbps	60	North America
Jupiter-2 (EchoStar XIX)		GEO	03/2017	Ka	220 Gbps	138	Americas
Jupiter-3 (EchoStar XXIV)		GEO	After 2020	Ka	500+ Gbps	301	Americas
Ka-Sat	Eutelsat	GEO	03/2010	Ka	90+ Gbps	82	Europe, Middle East (partly)
KONNECT		GEO	01/2020	Ka	75 Gbps	65	Europe and Africa
KONNECT VHTS		GEO	After 2020	Ka	500+ Gbps	230	EMEA
Intelsat Epic^{NG}	Intelsat	GEO	After 2020	C+Ku+Ka	25-60 Gbps	-	Global
Kacific1	Kacific	GEO	12/2019	Ka	60 Gbps	56	Asia Pacific
SES-12	SES	GEO	06/2018	Ku+Ka	40 Gbps	72	Asia-Pacific & Middle East
SES-14		GEO	01/2018	C+Ku	30 Gbps	-	Americas
SES-15		GEO	05/2017	Ku	25 Gbps	-	Americas
SES-17		GEO	After 2020	Ka	-	200	Americas
O3b mPOWER		MEO	Start from 2019	Ka	10 Tbps	30000	Global (50°N ∼50°S)
HYLAS 3	Avanti	GEO	06/2019	Ka	9 Gbps	8	EMA and Asia (partly)
HYLAS 4		GEO	04/2018	Ka	75∼100 Gbps	64	EMEA
Sky Muster™	NBN	GEO	2016	Ka	135 Gbps	101	Australia
OneWeb	OneWeb	LEO	Start from 2020	Ku	7+ Tbps	16 per satellite	Global
Starlink	SpaceX	LEO	Start from 2019	Ku	10+ Tbps	-	Global
Telesat	Telesat	LEO	2022 (planned)	Ka	10+ Tbps	1872	Global

Table 1 lists some of the representative commercial HTS systems and the key features, including the orbit, time of launching, frequency band, total throughput, number of spot- beams and coverage. We can find that the next-generation HTS using more spot-beams can achieve total throughput from tens of Gbps to as high as 10 Tbps, much higher than the current HTS systems. Besides, according to [9], the increase in the throughput of HTS leads to remarkably cheaper services, which is critical for the broad adoption and competitiveness of HTS services.

From Table 1, a majority of the state-of-the-art HTS systems use Ka band, which is between 26.5 GHz-40 GHz. Compared to the lower bands, Ka band offers a higher bandwidth for communications. Furthermore, Ka band enables implementing antenna arrays with large number of antennas, since the wavelengths are usually close to or smaller than 1 cm, which results in narrower spot beams with higher gains. However, compared to the lower frequency bands, the free-space loss is higher for the Ka band.

1.3 Frequency Reuse Techniques

HTS systems use spot beams in high frequency bands (e.g., Ka band), which enables the reuse of the frequency spectrum resources, resulting in increased system throughput. Fig. 1 shows the spot-beams used by Sky Mesh to provide broadband internet access to its customers in Australia [10]. Narrow beams with radius 125 km are used to cover coastal cities which have higher population densities, and small islands. Wide beams with radius 325 km are used to provide coverage for areas with low population densities, which is the case for central, Northern and North-Western Australia.

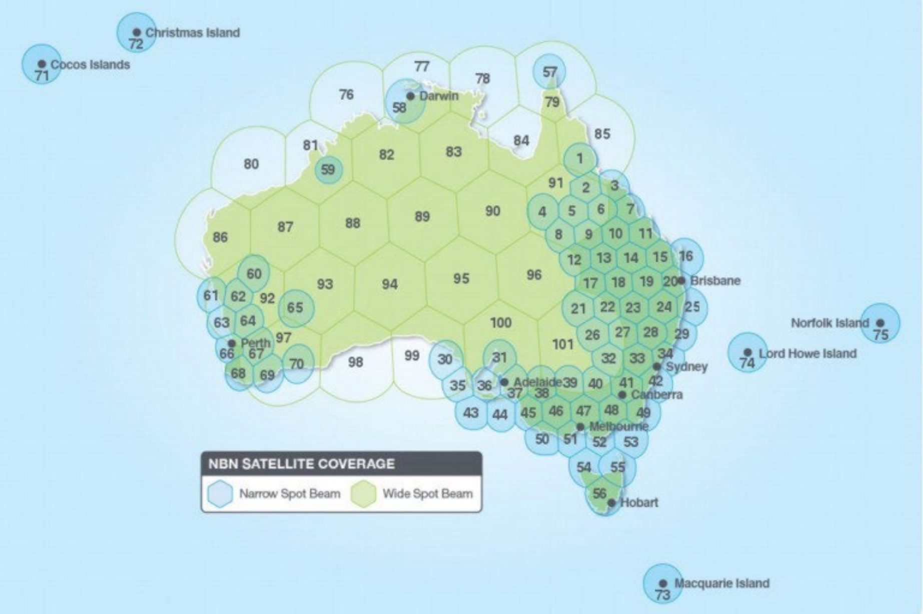
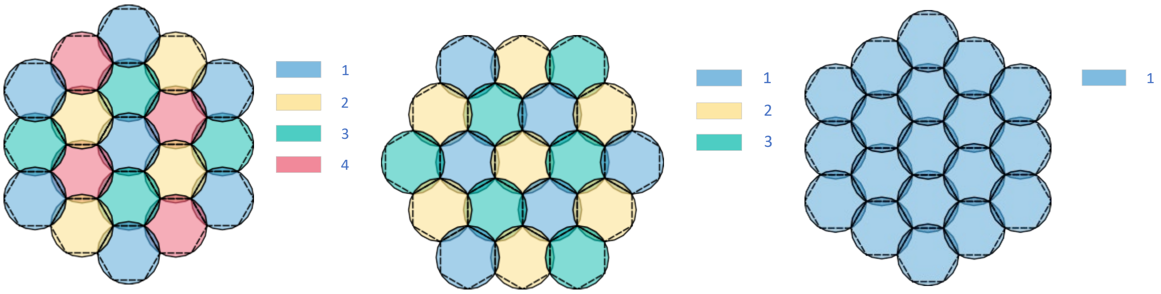


Figure 1: Spot beams used by Sky Mesh to provide internet access to its customers in Australia [10].

Frequency reuse, which allows the efficient use of frequencies many times on the same satellite, is a key feature of HTS generating multiple spot-beams. As shown in Fig. 4, there are different schemes for frequency reuse. The four-cell frequency reuse scheme shown in Fig. 2(a) is popular due to its interference isolation properties, which is typically achieved using two frequency bands and two polarizations in satellites.

A larger frequency reuse factor (i.e., the number of frequency reuse colours) means the whole frequency band is segregated into fewer sub-bands. HTS with smaller frequency reuse factors can achieve larger bandwidth and hence higher data rate. The most ag- gressive choice is the universal frequency reuse (UFR) scheme (Fig. 2(c)), which allows the satellite to maximize the available frequency band. However, the higher frequency reuse level also leads to more interference between beams, which can cause performance degradation of HTS systems.



(a) Four-color frequency reuse (4CFR) scheme.

(b) Three-color frequency reuse (3CFR) scheme.

(c) Universal frequency reuse (UFR) scheme.

Figure 2: Some frequency reuse schemes for spot-beam HTS systems.

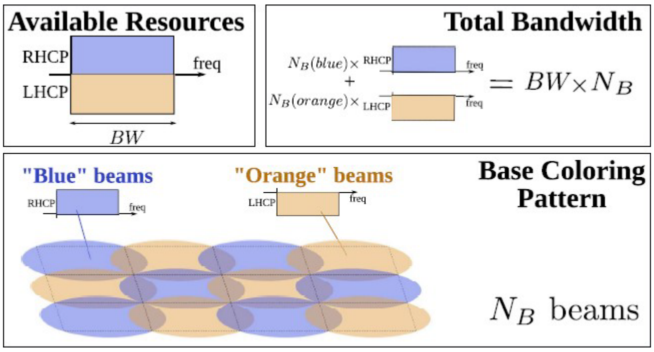
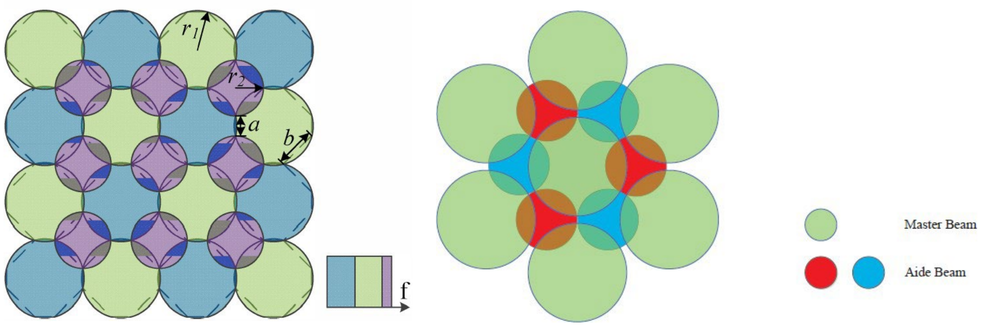


Figure 3: The two-color frequency-reuse scheme considered in [11].

The two-cell frequency reuse scheme shown in Fig. 3 (one frequency band and two polarizations) was considered in [11] for the spot-beam return link. A coordinated mod- ulation and coding selection process was proposed to mitigate the ensuing interference in the return link. Note that unlike traditional hexagonal beam arrangement shown in Fig. 4, the beams in [11] are arranged as a chess board.

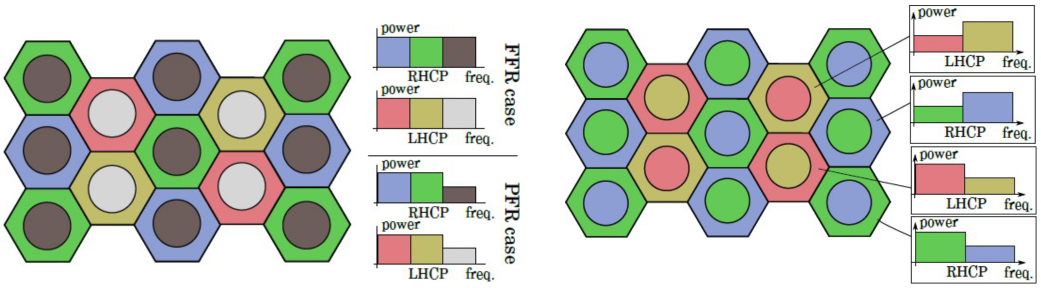
In [12], a dual-sized interleaved frequency reuse scheme for spot beams was proposed, which is shown in Fig. 4(a). The frequency spectrum was allocated across two large beams and one small beam. [13] also presented a dual-sized interleaved frequency reuse scheme for spot beams, which is shown in Fig. 4(b). It consisted of master beams and aid beams. The performances of [12] and [13] were evaluated in terms of frequency reuse factor, uplink interference and throughput per beam unit.

In [14], the performance of fractional fractional frequency reuse, partial frequency reuse (Fig. 5(a)) and soft frequency reuse (Fig. 5(b)) for multi-beam satellite return links



(a) The frequency reuse scheme proposed in [12]. (b) The frequency reuse scheme proposed in [13].

Figure 4: Dual interleaved frequency reuse schemes proposed in the literature.



(a) Fractional frequency reuse and partial frequency reuse [14]. (b) Soft frequency reuse [14].

Figure 5: The fractional frequency reuse schemes considered in [14]

is studied. The maximum throughput and proportional-fair criteria are considered as performance evaluation metrics for the fractional frequency reuse schemes.

Calculations to determine SINR as well as the (interference free) forward and return link budgets, and further analysis, for a spot beam system with four-cell frequency reuse are provided in [15]. In particular, see Tables III and V in [15] for results using typical satellite parameters. Results for SINR are given in terms of CDFs (coverage for a given threshold). From this and known modulation and coding schemes, throughputs can be determined.

1.4 Spot-Beam Antenna Technologies

The conventional approach for implementation of spot-beams is using multi-horn reflector antennas. A detailed review of multi-horn reflector antenna design technologies for spot- beam satellites can be found in [16]. The spot beams shown in Fig. 1 is an example of spot beams generated by multi-horn reflector antennas.

While the multi-horn reflector antennas are cost effective and has a matured tech- nology, it is not possible to reconfigure beams generated by multi-horn reflector antenna once the satellite is deployed, i.e. the coverage locations and the area covered by each beam remains fixed through out the life time of the satellite. Hence, it is difficult to efficiently accommodate temporary increases in demand in traffic in a small geographical area which occur after the deployment of the satellite (e.g.: the establishment of a mining town in Central Australia, a sports event attended by a large number of participants and spectators). Furthermore, fixed spot beams is not an efficient method to serve a moving object with large number of users with traffic demand (e.g.: a cruise ship, a train or an aircraft).

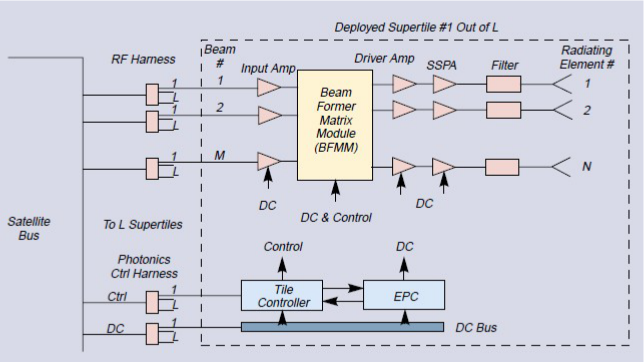


Figure 6: A typical block diagram of an active phased array antenna [17].

Direct radiating Active phased array (APA) antennas can be used to improve the flexibility in beam configurations. A typical block diagram of an APA antenna is shown in Fig. 6. [17] discusses the building blocks in APAs and the technologies used in APAs in detail. APAs can be used to dynamically synthesize beams of different shapes and sizes to cover a geographical location with traffic demand. Therefore, APAs can be used to adjust coverage based on the variations in demand after the satellite is deployed.

Direct radiating APAs also provide following additional advantages compared to their conventional counterparts. [17, 18].

1. Direct radiating APAs can radiate their waste heat into space, reducing the heat dissipation into spacecraft bus, which enhances the on-board power output. This will increase the throughput of the HTS system.

2. They provide larger apertures, which result in a smaller minimum spot size. Hence, APAs can be used to generate narrow beams to efficiently serve small towns, cruise ships, trains or aircrafts. Furthermore, these beams can be moved with the moving vehicle through adding a phase shift at each antenna.

The advancements in the Monolithic microwave integrated circuit (MMIC) technology has enabled the creation of large APA antenna arrays as flat-panel designs, which are both thin and have low mass. In [19], the design of a planar APA system with 256 antennas for the Ka band was presented. Each RF chipset was developed using MMIC technology, to which the antennas and the cooling plate were attached at the top and the bottom, respectively.

Amongst the technologies considered in [18] are the deployment of active phased ar- rays, which can now have much larger surface area as they are deployed in orbit much in the same way as solar panels, as shown in Fig. 7.

1.5 Planning and Designing of Spot-Beam Systems

An algorithm for planning spot-beam satellite systems considering both technical and commercial requirements was proposed in [20]. The technical considerations included payload power requirements, antenna aperture and available bandwidth for gateways

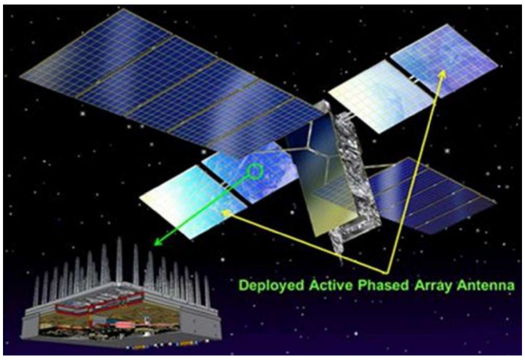


Figure 7: An illustration of the deployment of active phased array antennas in a satellite [18]

and user terminals, and the commercial considerations included budget of deployment, quality of service and minimum required service availability. The four-color frequency reuse scheme with two frequency bands and two polarizations was used. Multi-horn reflector antennas were used for generating spot beams.

In the proposed algorithm, system capacity was maximized with respect to four design parameters, which are service area filling percentage, the availability percentage of the satellite link, the cost of satellite and the mass of the satellite. Simulation results were presented for the coverage of France and Central Western Africa.

A feasibility study for a terabit/s satellite system with multiple gateways using the DVB-S2 and DVB-RCS2 adaptive coding and modulation schemes was provided in [21]. In addition, various frequency reuse schemes were investigated. The authors considered a reflector antenna at the satellite. Ka band was used for the user links and Q/V bands were used for the feeder link. Various performance results were presented including SINR cumulative distribution functions (CDFs), offered capacity for different beam cases and probability density functions (PDFs) for the modulation and coding schemes. Subse- quently the question of High Powered Amplifier (HPA) sharing was considered and the Total Power Requirement was determined for various implementations (in terms of fre- quency reuse). With four-color frequency reuse, it was shown that the performance of the forward link is limited by noise and the performance of return link is limited by interference.

In [18], the authors examined various new enabling technologies for high throughput satellite (HTS) communications. Figure 3 of [18] tables the value propositions: market viability, capacity, traffic, QoS, future proofing etc. Table II looks at various technologies and their benefits to these value propositions. As one example, in-orbit reprogrammable processing addresses the lag in technology at the time of launch and during the life time of the satellite, which can be more than 15 years.

The authors also explored payload processing which has been implemented using highly customised hardware. Here, the principal consideration is SWaP (Size, Weight and Power). Depending on the implementation, there is a tradeoff between the degree of flexible reprogramming on the one hand and SWaP at the other, see figure 8. For high power missions the authors suggest that 22kW of power can be achieved whilst having low solar array mass as an evolution from existing technologies.

1.6 Key Findings and Future Research

- It was shown in [21] that in the forward link, the four-color frequency reuse scheme is mainly noise limited. However, the system becomes interference limited with more aggressive reuse schemes (e.g., two-color reuse or UFR). More aggressive reuse techniques are necessary to provide broadband services to large number of users using limited frequency spectrum resources. Therefore, interference mitigation is imperative for spot-beam HTS systems.

- Conventional fixed spot-beam satellites may radiate on-board transmit power, which is a limited resource, to beams with zero traffic demand (dark beams). This is clearly a waste of valuable resources. Adaptive resource allocation strategies are required to be implemented at the satellite to allocate resources based on traffic demand. The research on such strategies are still in their infancy and consider only very simple system models and performance metrics, which are inadequate for practical systems. More sophisticated system models and performance metrics need to be studied for adaptive resource allocation based on traffic demand.

- Direct radiating active phased arrays (APAs) are highly beneficial in terms of flexibility and efficiency in HTS systems. However, for narrow beams in the order of few kilometers (e.g., a small mining city), the beam width has to be a small fraction of a degree at the GEO orbit. This requires APAs with very large number of antennas (in the order of 1000s). Due to the smaller wavelength of Ka band, this is possible to achieve with planar arrays which span several meters. However, wider implica- tions of the deployment of such large Ka-band APA arrays at the satellites in terms of heat generation, RF interference, weight, cost, life time of the communication payload and so on, need to be investigated.

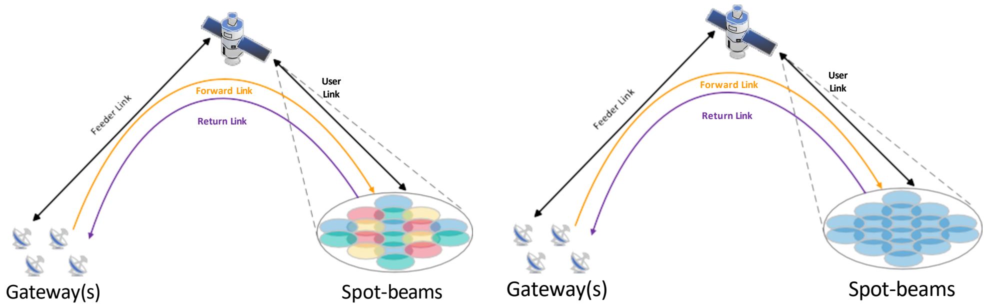
- Serving a large number of customers in a large moving vehicle (e.g., a cruise ship, a train, an aircarft) efficiently is an important application of APAs. In such appli- cations, the beam generated by the APA has to follow the vehicle. Such a system has not been studied or implemented yet. However, the potential benefits of such a system makes a compelling case for further research in this area.

- The high capacity requirements of HTS also give rise to bandwidth and capacity requirements to the feeder link. Therefore, gateway diversity/ multiplexing schemes and even higher frequency bands (e.g. Q/V bands) may have to be used for the feeder link.

Advanced signal processing techniques provide powerful solutions to some of the above challenges and have received strong interests from both academia and industry. In the next sections, we will discuss precoding and user scheduling techniques for HTS in GEO.

2 Multibeam Precoding and User Scheduling for HTS

Typical 4CFR and UFR HTS system architectures are shown in Fig. 8. The links from the satellite to the UT, and from the gateway(s) to the satellite are called the user link and the feeder link, respectively. Up to the direction of communications, the links from the gateway(s) to the UT and from the UT to the gateway(s) are known as forward link and return link, respectively.



(a) A HTS system architecture using 4CFR scheme.

(b) A HTS system architecture using UFR scheme.

Figure 8: HTS system architectures of HTS using typical frequency reuse schemes.

Although larger frequency reuse factor can mitigate the interference between spot- beams, it also directly limits the available bandwidth of the HTS systems, and may reduce the capacity of communications. Therefore, implementing aggressive UFR schemes that allow the use of the whole frequency band becomes the most efficient way to improve the total throughput. With the use of aggressive UFR, interference mitigation techniques be- come mandatory as adjacent beams suffer from severe multiuser interference. Precoding techniques [22–27] have excellent performance in mitigating the interference and control- ling the transmitted power. Therefore, efficient multibeam precoding design is desired for the next generation HTS.

The channel state information (CSI) is essential in performing precoding. Due to the long distance between the satellite and user terminals (UTs) and the line-of-sight (LoS) channel, the UTs served by one beam generally have similar amplitudes of channel vectors. However, the phases of the channel vectors are usually different phase effects generated by the feeder link, local oscillators (LOs) on the satellite board, and the LOs of the user terminal [22]. Therefore, the phase differences of channel vectors among UTs are not negligible. Due to the multicast nature of HTS where more than one UT is often addressed by one frame, if the users belonging to the same beam have significantly different channel vectors, the precoding performance will deteriorate. Through user scheduling (also called clustering or grouping) techniques, the HTS system can smartly group users with similar channel vectors. Hence, user scheduling is also a crucial technology that can maximize the objective of interest and improve the performance of HTS.

2.1 Precoding and User Scheduling in HTS and Cellular Networks-Similarities and Differences

Since precoding and user scheduling technology has been widely studied in terrestrial wireless communications, in this section, we will compare the similarities and differences of this technology in HTS and cellular networks.

The similarities of precoding and user scheduling technology are summarized as follows [28]:

- Multibeam HTS systems presents certain similarities to the cloud radio access net- work (C-RAN) architecture, where the baseband processing and the RF elements are placed in separate locations. For example, the scenario where the HTS system has multiple gateways (GWs) shares similarities with the multicell C-RAN systems, and the onboard processing in HTS systems performs like relay processing in terrestrial networks.

- The signal processing approaches in multibeam HTS are similar to the multigroup multicast strategies in terrestrial wireless communications.
- The hybrid on-ground/onboard precoding, which can reduce the bandwidth requirements in the feeder link, presents certain similarities to the hybrid digital-analog precoding techniques in terrestrial mmWave systems.

Despite the above similarities, the precoding and scheduling techniques in HTS have significant differences, as summarized below:

- Different to the antennas of small size in cellular networks, HTS systems have extremely large-scale dimensions of antennas with high directivity, to improve the SINR of signals.
- Because there are usually few scatterers near the satellite in the space, the channel is often characterized by an LoS path. Therefore, the conventional spatial diversity enjoyed by terrestrial cellular networks, e.g., massive MIMO, is no longer applicable to HTS systems.
- The limited power consumption, hardware complexity, and signal processing capability of the satellite board prompt new signal processing strategies, considering unique optimization objectives and constraints. For example, instead of the common sum-power constraint, precoding in HTS usually prefer the per-antenna power constraints to satisfy the power limitation on each feed.
- The channel modelling of HTS can be influenced by the atmosphere, as well as the significant delay caused by the long-distance propagation. Besides, users served by the same spot-beam have more similar channels than the ones of cellular networks. For example, for different users, the amplitude variations of the channel vector are much smaller than their counterparts in cellular networks.

We now go on to discuss various proposals for precoding and user scheduling in the literature. As will be seen, there are already many diverse proposals for these in the literature. However, there are no ‘off-the peg’ solutions. And, no matter which approach is adopted, extensive investigations will be needed in order to justify their implementation.

2.2 Joint Precoding and User Scheduling

Assuming perfect knowledge of CSI at the transmitter, researchers with University of Luxembourg developed precoding methods for GEO satellites [23]. They proposed the linear multicast aware minimal means square error (MMSE) method based on [29], and the multigroup multicast beamforming under per-antenna transmit power constraints (also called as weighted max-min fair precoder), developed from their earlier work in [30]. It is shown that when two users are assumed per frame, the MMSE and weighted max- min fair precoders can respectively achieve 21% and 42% gains of per beam achievable throughput, compared with the conventional 4CFR scheme. Although the throughput could decrease if the number of users per frame increases, the weighted fair precoder was shown to outperform the conventional 4CFR scheme. However, the methods in [23] is quite heuristic with ideal assumptions, such as the perfect CSI, and the ignorance of the interference from adjacent clusters. Moreover, no user scheduling was considered in this work.

As we mentioned before, user scheduling that can group users with similar channel vectors has shown great potential in improving the throughput performance of the HTS. The authors in [24] heuristically investigated user scheduling based on the geographical location of the users. In this work, it is assumed that geographically closed users have similar channel matrix so that encoding them with the same frame can improve the pre- coding performance. A linear MMSE precoding approach developed in [31] was applied. It was shown that with perfect channel state information at the receiver (CSIR) and in- stantaneous channel state information at the transmitter (CSIT), the precoding and user scheduling method can achieve gains of total average throughput from 104% to 53% (with 2 to 10 users per group), compared with the conventional 4CFR scheme without precod- ing. Relaxing the perfect assumptions and considering the imperfect CSIR and outdated CSIT, it can still achieve 44% to 3% (with 2 to 10 users per group) of throughput gain.

Later, the same authors of [23] proposed more precoding and scheduling approaches in [25]. Instead of the max-min fair optimization, [25] considered the per-antenna power- constrained sum-rate maximization objective (called SR precoder [25]), and also studied the additional constraints on the minimum rate (called SRA precoder in [25]). Taking the modulation defined by DVB-S2X into account, they further developed the modulation aware max SR optimization (called SRM precoder in [25]). A multicast-aware scheduling algorithm based on the semi-orthogonality criteria [32] was also proposed and can further improve the precoding performance. The simulation results in [25] show that the SR, SRA, and SRM precoders can achieve noteworthy gains of the per beam throughput over the max-min fair solutions and conventional 4CFR scheme (refer to Table 4). Besides, the proposed SRM precoder with scheduling can obtain approximately 20% to 49% of throughput improvement to the SRM precoder with random scheduling. However, the precoders solved nonconvex optimization problems using semi-definite relaxation (SDR) and Gaussian randomization, which can lead to immense computational complexity with a large number of beams in the next-generation HTS systems.

In order to reduce the computational complexity of precoding algorithms in HTS systems, a new two-stage precoding strategy by dividing the precoding matrix into two multiplicative sub-matrices was proposed in [26]. Firstly mitigating the inter-beam in- terference and then optimizing the intra-frame data rate, the proposed linear precoder can largely decrease the computational complexity and achieve improved precoding per- formance. In [26], the authors also developed a k-user grouping approach where not only the amplitudes but also the phases of the channel vectors are taken into account for the user scheduling. When 3 users are served in one beam, the proposed multibeam inter- ference mitigation (MBIM) precoder can achieve a throughput increase of at least 1.5% with respect to the SRM scheme while offering a substantial computational complexity reduction. With user scheduling, the MBIM can further increase the average throughput per beam at least 12%.

Recently, an interesting joint scheduling and precoding approach [27] has been de- veloped based on the SRM precoder in [25]. The binary variable constraints reflecting the scheduling choices were formulated in the SRM optimization problem. This problem was reformulated to a difference-of-convex/concave (DC) problem and iteratively solved through the convex-concave procedure (CCP). Compared with the SRM precoder, the proposed pre-selection based joint scheduling and precoding method achieves 15% to over 50% gain of the average throughput per beam, with the ratio of the number of pre-selected users to the users per frame varying from 4 to 1.

A geographical user scheduling algorithm (GSA) for multicast precoding in the forward link of a fixed spot-beam satellite system was proposed in [33]. In the proposed GSA, each beam coverage area is divided into sectors (scheduling sectors) and at a given time, the scheduler only schedules users located in the same scheduling sectors in each beam. This mitigates the loss in performance of precoding due to channel correlations exist between users who are located close-by and covered by different beams.

2.3 Robust Precoding in Multibeam HTS

Inaccurate and outdated CSI can largely influence the performance of the above precoding and scheduling approaches. The literature [34] studied the precoding and scheduling with delay CSI in L band mobile interactive satellite services. Channel models under various mobile scenarios, including slow nomadic, maritime, and maritime low elevation were consolidated. In the scheduling algorithm proposed in [34], for each beam in each time step, a user is first randomly selected, and then the remaining users to be scheduled are selected to be the set of $N_u - 1$ which has channel coefficients closest to the channel coefficient of the first user in terms of Euclidean distance, where N_u is the number of multicast users scheduled per beam. Clearly, the performance degrades as N_u increases. For all the studied scenarios, compared with the conventional 4CFR scheme, 20% and 40% throughput gains were achieved with 5 and 2 users per frame, respectively.

Considering the perturbation of the channel matrix, a robust inter-beam precoding method was proposed in [26]. A robust k user grouping algorithm was also proposed in [26], which considered the uncertainties in the CSI estimates, where the CSI uncertainties were represented by an additional scalar parameter added to the original k user grouping optimization problem considered in the same paper. When only imperfect CSI is available, the robust MBIM precoder can achieve 3% throughput gain than the MBIM precoder mentioned before with 3 users per beam.

In [35–37], robust precoding methods considering the channel phase uncertainty due to the outdated CSI were developed under the FFR scheme. The common underlying intuition for these works is to model the phase uncertainty as a random process and introduce the outage constraint to ensure that the outage probability is maintained at desired levels. Different optimization objectives are studied in these works, including minimizing the lowest transmitted power from each feed [35], the max-min fair SINR [36], and the sum-rate maximization [37]. The simulation results showed that when the non- outage probability threshold equals to 0.8, almost 50% of the sum rate gain can be achieved by the robust precoder in [36], compared with the conventional approach which adopts the outdated CSI directly as the true CSI. The robust precoding method proposed in [37] can further improve the performance.

2.4 Precoding under Limited Feeder Link Bandwidth

In HTS systems, the feeder link is required to support the overall satellite traffic and perform precoding. Therefore, there is an increasing bandwidth requirement of the feeder link, which can increase the frequency spectrum congestion. Three types of solutions have been proposed towards this problem, as listed below:

1. Moving the feeder link to the higher frequency, such as Q/V/W bands [38, 39] and optical frequencies [40, 41].
2. Employing multiple gateways (GWs) [26, 42–44].
3. Implementing onboard signal processing [45, 46].

Challenges of moving the feeder link to higher frequencies are mainly due to the prop- agation attenuation and atmospheric effects. Moreover, feeder links using Q/V/W bands can cause potential interference with other microwave systems. In addition, when the number of beams is very large, the bandwidth requirement can still be unaffordable even the higher frequency feeder links are employed. Since the key techniques of moving the feeder link to higher frequencies are less related to precoding, we will focus on precoding techniques employing multiple GWs and onboard signal processing .

2.4.1 SatCom with Multiple GWs

For the SatCom industry, there is a trend that a larger number of GWs with smaller size will be implemented in the next-generation SatCom systems. For example, the following Table 2 summarizes the number and size of gateways for ViaSat 1-3 [47].

Table 2: A comparison of number and size of GWs for ViaSat systems.

Systems	ViaSat-1	ViaSat-2	ViaSat-3
Size of GW dish	7 m	About 4 m	Less than 2 m
Number of GWs	20	45	hundreds

Similarly, the non-geostationary (NGSO) systems such as Telesat, OneWeb, and SpaceX also considers large numbers of GWs. Literature [48] compared and simulated these three systems, the results related to GWs are listed in the following table.

Table 3: A comparison of GWs for the next-generation NGSO systems [48].

Systems	Telesat	OneWeb	SpaceX
Size of GW dish	3.5 m	2.4 m	5 m
Number of ground locations for maximum throughput	42	71	123
Required number of GWs per ground station	5-6	11	30

As pointed in [49], the major cost of future HTS systems will be absorbed by the ground segment development and management. Therefore, it is necessary to study the implementation of multiple GW network. The benefits of employing multibeam precoding over multiple GWs are as follows [28, 42, 47]:

- Using more GWs with the smaller size can significantly reduce the cost of the ground infrastructure.
- Multiple GWs can reduce the multibeam signal processing complexity, since each GW has to handle a smaller number of beams, compared with the single GW scheme.

- Aggregating the bandwidth of the feeder links of different GWs can alleviate the bandwidth requirement of the feeder link.

- In case one of the GWs fails or has very adverse fading, the traffic can be rerouted through other GWs to avoid service outage.

There are two types of issues to be solved for implementing multiple-GWs. The first issue lies in the placement and network architecture of GWs, and the second one is related to the precoding and user scheduling design in multiple GW networks. Both of the issues have a significant impact on satellite throughput and QoS.

I. Design of the architecture of gateway networks

The design of the architecture of the gateway network is flexible with some restrictions. As mentioned in [50], the gateways do not have to be distributed globally to provide global coverage. Conventionally, the placement of terrestrial gateways has been studied considering the following practical aspects [50]:

- Geographical, topographic and meteorological factors: No obstructions, mild temperatures with a very dry climate, the absence of common natural disasters, ample land to install devices and expand in the future, etc.

- The scale and size of gateway stations.

- Economic considerations: Electrical supply, cost of implementation and maintenance, proximity to a talented technical labour pool, neighbouring developments (can obstructing line of sight to the satellites), and access to fiber providers to upgrade the networks and keep up with regular maintenance.

- Political considerations: borders and ownership of the land, civil unrest or war zones, and religious impacts.

Recently, studies on gateway placement [49, 51–54] take more communication perfor- mance metrics into consideration, including but not limited to the overall cost, traffic routing, reliability, target throughput, propagation latency, and power consumption etc. The GW placement in novel scenarios such as integrated satellite-terrestrial networks (ISTNs), large-scale constellation networks with inter-satellite links is investigated by literature [51–53] and [54], respectively.

Study [49] funded by INMARSAT Ltd. and European Space Agency designs a strategy to group the GW within clusters of GWs. The proposed approach can select the best GW configuration given a set of practical constraints, such as possible GWs locations, characteristics of satellite/ground terminals, target system availability, so as to reach target system performance. The algorithms have been tested in a practical HTS system scenario and serve a good reference for the optimization of gateway architecture in the real world.

Several works [51–53] investigate the joint controller and gateway placement in ISTNs with GEO satellites. The gateway placement problem is usually modelled as a combinato- rial optimization problem and solved by heuristic algorithms. Based on software-defined network (SDN), literature [51, 52] proposed simulated annealing-based algorithms max- imizing the network reliability with low latency. Research [53] proposed joint satellite gateway placement and routing for ISTNs minimizing the overall cost of gateway de- ployment and traffic routing, with constraints on average delay requirement for traffic demands. Simulation results in these works show that the smart design of GW topology brings many benefits on the QoS of the ISTNs networks.

In NGSO systems, the gateway placement should also take account of the movement of the satellites and the handover process. There can also be satellite gateways that establish feeder links in space. The authors in [54] investigated the multiple gateway placement in LEO systems with inter-satellite links. A genetic algorithm (GA)-based method is proposed aiming at achieving optimal overall performance including delay, traffic peak, and load balance, with constraints on potential gateway location, gateway- satellite connectivity, and max hop-count. However, this work uses a simplified model, which divides the earth surface into discrete grid cells, without considerations on the practical issues.

In practice, it is a complicated problem with multiple interacted factors from diverse aspects. Gateway placement also directly influences the performance of precoding ap- proaches, since the precoding process is mostly performed at GWs. However, this problem has not been studied yet, and it is still unclear how the architecture of GWs can influence the performance of precoding.

II. Precoding and user scheduling under multiple GWs

With multiple GWs, there comes a few changes for precoding and user scheduling [28]. Firstly, since each GW can use only a subset of feed signals for performing the interference mitigation, the available degrees of freedom of precoding is reduced. Secondly, each GW can only have access to the CSI of its served users, whereas the CSI for the adjacent beams are still needed to reduce the interference. Therefore, information sharing between GWs are required, which can cause extra communication overload. Thirdly, perfect connectivity between GWs might not be possible in real deployments.

The concept of cooperating multiple GWs in FFR satellite communications was firstly proposed in [42]. Sum-rate maximization methods with a constraint on the transmitted power for each gateway were proposed, considering the partial CSI and data sharing respectively. The results in [42] show that using extra CSI and data sharing among non-overlapping GWs gives about 2% and 15% higher throughput than the one without cooperation between GWs respectively, which is about 28% and 42% larger than the throughput achieved by the conventional 4CFR scheme.

A centralized max-min fair precoding where each GW share its local CSI among all the other GWs was proposed in [43]. Due to the exchanging of CSI between multiple GWs, the proposed method achieves a higher minimum SINR, than the multi-GW precoding without any coordination. However, the information sharing among all the GWs is costly and hence quite impractical.

In [26, 44], different scenarios with different level of cooperation between GWs are comprehensively studied, including Individual Cluster Multibeam processing (ICM), 4 Gateways Collaboration (4GC), 7 Gateways Collaboration (7GC), Gateway Collabora- tion Multibeam processing (GCM), and Limited Multi-gateway Collaboration processing (LMC). The single gateway scenario is also examined as the reference (Ref) scenario. The average throughput can be found in the following Fig. 10. It is shown that the CSI sharing among adjacent clusters, as well as the LMC method which collaborates with all the gateways by transmitting the rank one approximation of their channels achieve good trade-off between gateway cooperation overhead and overall system performance.

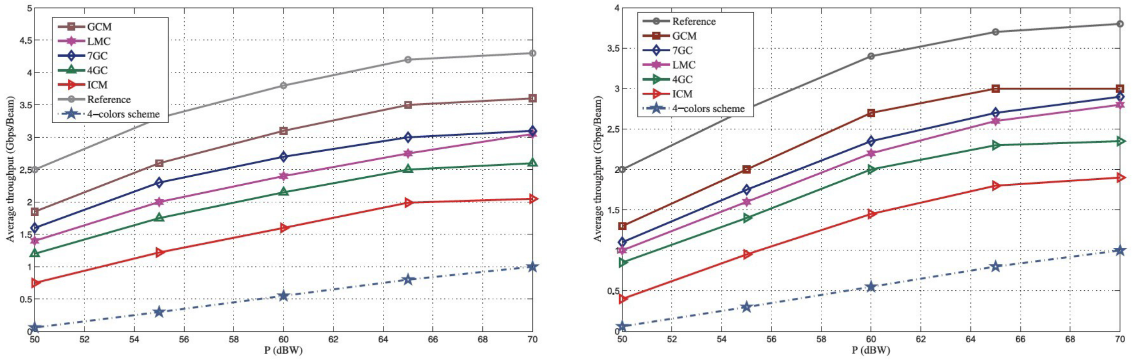


Figure 9: Average throughput considering multigateway block regularized precoding and different collaborative architectures. The intra-cluster interference is mitigated via MMSE (Left) ZF precoding (Right) [44].

2.4.2 On-board Signal Processing

There are several issues to be solved for the multiple-GW precoding [28]. Firstly, since each GW can use only a subset of feed signals for performing the interference mitigation, the available degrees of freedom of precoding is reduced. Secondly, each GW can only have access to the CSI of its served users, whereas the CSI for the adjacent beams are still needed to reduce the interference. Therefore, information sharing between GWs are required, which can cause extra communication overload. Thirdly, perfect connectivity between GWs might not be possible in real deployments.

The concept of cooperating multiple GWs in FFR satellite communications was firstly proposed in [42]. Sum-rate maximization methods with a constraint on the transmitted power for each gateway were proposed, considering the partial CSI and data sharing respectively. The results in [42] show that using extra CSI and data sharing among non-overlapping GWs gives about 2% and 15% higher throughput than the one without cooperation between GWs respectively, which is about 28% and 42% larger than the throughput achieved by the conventional 4CFR scheme.

A centralized max-min fair precoding where each GW share its local CSI among all the other GWs was proposed in [43]. Due to the exchanging of CSI between multiple GWs, the proposed method achieves a higher minimum SINR, than the multi-GW precoding without any coordination. However, the information sharing among all the GWs is costly and hence quite impractical.

In [26, 44], different scenarios with different level of cooperation between GWs are comprehensively studied, including Individual Cluster Multibeam processing (ICM), 4 Gateways Collaboration (4GC), 7 Gateways Collaboration (7GC), Gateway Collabora- tion Multibeam processing (GCM), and Limited Multi-gateway Collaboration processing (LMC). The single gateway scenario is also examined as the reference (Ref) scenario. The average throughput can be found in the following Fig. 10. It is shown that the CSI sharing among adjacent clusters, as well as the LMC method which collaborates with all the gateways by transmitting the rank one approximation of their channels achieve good trade-off between gateway cooperation overhead and overall system performance.

Other than the above techniques using multiple GWs, the onboard signal process- ing [45, 46] can also relieve the bandwidth requirements of the feeder link. A feed se- lection based hybrid on-ground/onboard precoding concept was proposed in [45] for the forward link of multibeam mobile satellite systems. The results given in this article show

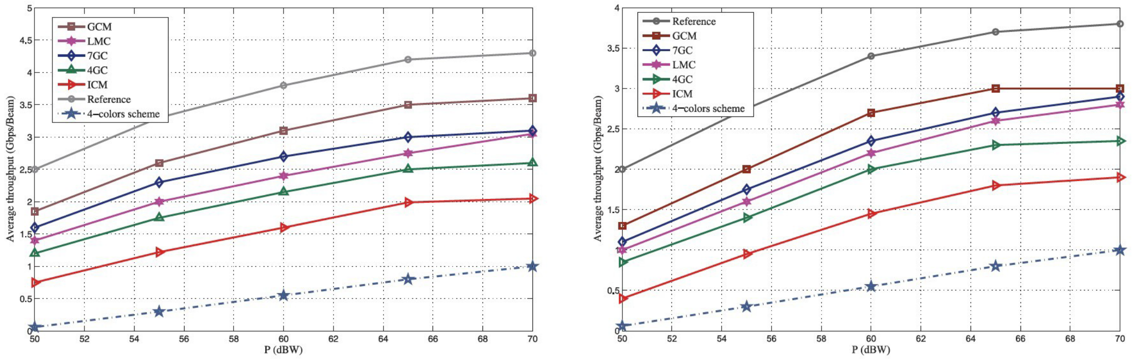


Figure 10: Average throughput considering multigateway block regularized precoding and different collaborative architectures. The intra-cluster interference is mitigated via MMSE (Left) ZF precoding (Right) [44].

that the proposed approach can bring a significant complexity reduction with affordable performance degradation. However, the S-band considered in their work is less practical for the next generation HTS systems.

Authors in [46] developed a new hybrid onboard on-ground multibeam satellite architecture, as well as a robust linear MMSE precoding and detection approach technique in both the forward and return link. Results in [46] show that in some scenarios, the approach can increase the spectral efficiency over the 6% and 15% for return and for- ward links, respectively. However, no user scheduling technique was considered in hybrid on-ground/onboard precoding. More research is expected in the future.

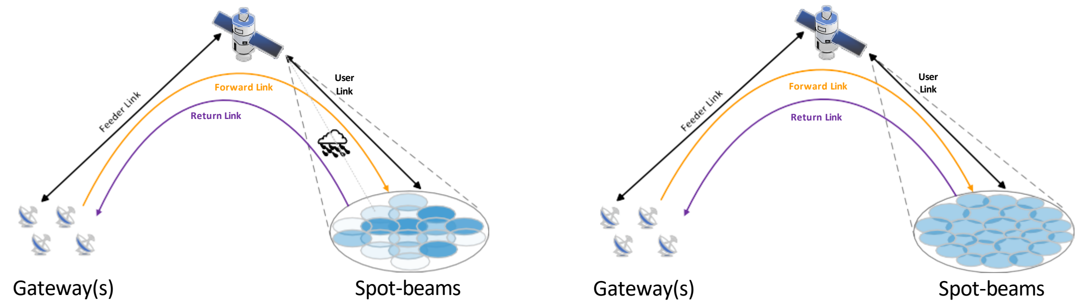
2.5 Adaptive Precoding and User Scheduling Based on Traffic Demands

Adaptive multibeam precoding and user scheduling techniques, providing significant flex- ibility in resource allocation and optimization, are also promising for future HTS. With such techniques considering the limited resource and traffic demands, the performance of HTS can be dramatically improved in terms of mitigating interference, improving the robustness to rain fade, and saving resource, such as the power and bandwidth. In this domain, various adaptive techniques can be explored from the following aspects.

- One aspect is to adaptively adjust the precoding coefficients, such as the beamwidth, bandwidth, power, and coverage of the spot-beams, according to the channel con- dition, user requirements, and interference. As shown in Fig. 11, by using adaptive precoding, the achieved beamwidth and communication capacity can vary from beam to beam.

- Another aspect is to implement the adaptive user scheduling, which refers to the HTS can adaptively choose and group the users that can simultaneously receive the signals precoded by a set of precoding coefficients within the same beam. This is trying to explore the multi-user diversity or multiplexing gain to improve the throughput, reduce the interference, and make full use of limited resources.

So far, most of the studies focus on the dynamic resource allocation among spot-beams, where adjacent beams use different frequency band to mitigate the interference [55–60].



(a) HTS system adaptively achieving different capacities for different spot-beams. Some areas can be affected by rain attenuation.

(b) HTS adaptively with adaptive beamwidth and coverage range.

Figure 11: HTS system architectures with adaptive precoding.

Besides, there are also a few precoding designs based on traffic demands, considering full frequency reuse (FFR) scheme [61–63].

The earlier research [55, 57] studied power optimization methods for each beam based on traffic demands among different users. However, these primitive schemes ignore the in- terbeam interference, which is indeed non-negligible for practical multibeam HTS systems. Later, literature [58] further develops the power allocation scheme employing a two-stage optimization, considering both traffic match and power consumption mitigation. With considerations of the co-channel interference, numerous metaheuristic algorithms are used to solve the NP-hard optimization problem, under the four-color frequency reuse (4CFR) scheme. As shown in Fig. 12, the Genetic Algorithm - Simulated Annealing (GA-SA) algorithm proposed in [58] can meet the assumed traffic demand for different beams. How- ever, as can be observed in the 36th beam, when the demanding capacity is large, power allocation has limited capability. This phenomenon can also be found in the simulation results in [56, 59]. We reckon that it is due to the power limitation for each beam, and the logarithm operation of SINR when calculating the Shannon capacity.

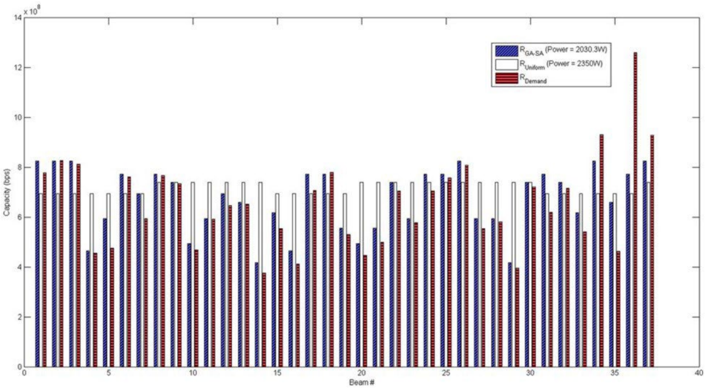


Figure 12: The capacity achieved by the GA-SA algorithm.

From another perspective, in [56, 59, 60], dynamic bandwidth are allocated to different beams to match the traffic demands. In literature [56], an iterative optimization algorithm to minimize the unmet traffic is studied. A more advanced objective function considering fairness is proposed in [59]. Using the seven-color frequency reuse (7CFR) scheme, [60] developed the unmet traffic minimization problem by addressing the interbeam interference and time delay. Jointly optimizing the bandwidth and power allocated to each beam, these methods are proved to be more efficient compared with the aforementioned power- only optimization schemes. However, the total throughput of such HTS systems can be limited by the available bandwidth of each spot-beam.

In a different context employing the full frequency reuse, adaptive precoding techniques are studied considering the beam power and traffic demands [61–63]. In these studies, precoding weight coefficients are designed in the scenario where only one user is served by one beam. In a more practical scenario, for the next-generation multibeam multicast HTS, usually multiple users are simultaneously served by one beam. Therefore, user scheduling is also an important technology which can smartly group the users to improve the precoding performance. Users in a scheduled/clustered group will receive the same signal frame precoded by the same precoding weight vector. To our best knowledge, the geographical position and channel state information (CSI) of different users are the primary factors considered by the existing user scheduling studies [25–27]. However, as we can see from the above discussions and simulation results in [61–63], the unbalanced traffic demand can largely influence the precoding performance in terms of the throughput.

Precoding and power allocation schemes only are insufficient to serve the extremely large traffic demands in a specific area. Therefore, user scheduling methods that jointly considering users' geographical position, CSI and traffic demands is worth researching. Moreover, user scheduling can further influence the precoding design constrained by both traffic demands and power consumption.

2.6 Potential Gains, Challenges, and the Industry Gap

Table 4 presents representative results from the studies in [23–27] which summarize the potential gains from precoding and user scheduling. In the table, we use "conventional 4CFR" to refer to the four-colour reused (4CFR) scheme without using either precoding or scheduling techniques. From the table, we can see that both precoding and scheduling are shown to have great potential in significantly improving the throughput of HTS.

Due to the peculiarities of SatComs, there are a number of challenges of the design of precoding and user scheduling in HTS systems, as listed below:

- Multicast/broadcast precoding is usually considered in multibeam HTS systems.
- Low SNR caused by the significant propagation attenuation in SatCom scenario.
- The control of total power consumption is insufficient, and it is critical to limit the power of each antenna feed on the satellite payload.
- Compared with cellular networks, there are more rigorous computational complexity requirements for the precoding and user scheduling algorithms, due to the remarkable latency, system limitation of satellites and large number of beams to be generated.
- Because of the LoS propagation, the spatial diversity is less applicable in the design of precoding and user scheduling approaches in HTS.
- In HTS, the acquired CSI is usually outdated and influenced by atmospheric effects, which also limits the performance of precoding and user scheduling.

Table 4: The gains of throughput achieved by representative precoding and scheduling techniques in HTS systems.

Literature	Methods	Frequency reuse scheme	Potential Throughput Gains	Attributes
[23]	MMSE	UFR	21% of gains over the conventional 4CFR with 2 users/frame	Precoding only
	Weighted max-min fair precoder		42% of gains over the conventional 4CFR with 2 users/frame	Precoding only
[24]	MMSE + Geographical Scheduling	4CFR	104% to 53% of gains over the conventional 4CFR with 2 to10 users/cluster (perfect CSI) 44% to 3% of gains over the conventional 4CFR with 2 to10 users/cluster (imperfect CSI)	Precoding + Scheduling
[25]	SR precoder	UFR	64% to 2% of gains over the conventional 4CFR with 2 to 6 users/frame	Precoding only
	SRA precoder		61% to -5% of gains over the conventional 4CFR with 2 to 6 users/frame	Precoding only
	SRM precoder		70% to 6% of gains over the conventional 4CFR with 2 to 6 users/frame	Precoding only
	SRM precoder + scheduling		20% to 49% of gains over the SRM precoder using random scheduling with 2 to 6 users/frame	Precoding + Scheduling
[26]	Two-stage MBIM precoder	UFR	At least 1.5% of gains over the SRM precoder with 3 users/beam	Precoding only
	Two-stage MBIM precoder + k-user grouping		At least 12% of gains over the MBIM precoder with 3 users/beam	Precoding + Scheduling
[27]	Joint scheduling and precoding SRM approach	UFR	15% to 50% of gains over the SRM precoder with the ratio of the number of pre-selected users to the users per frame varying from 4 to 1	Precoding + Scheduling

- There are hardware impairments in the satellite system, such as the nonlinearities of the HPA and the frequency/phase instabilities of LOs. These impairments have a nonnegligible impact on the CSI and can hence affect the precoding and scheduling performance.

- To support the overall satellite traffic and perform precoding, there arise more bandwidth and capacity requirements for the feeder link.

- The onboard signal processing capability is highly restricted by the satellite platform and payload.

- Considering the SatCom standards, such as DVB-S, DVB-S2, DVB-S2X, 3GPP TR 22.822/23.737/ 38.811/38.821 that provide standardization of signals, such as frame structure and modulation schemes, can also bring benefits to the design of precoding and user scheduling schemes.

- To improve the quality of service, novel precoding approaches considering the traffic demand in different areas are desired.

When designing and implementing precoding and user scheduling for HTS, the above challenges are expected to be addressed.

Although precoding for high throughput satellite communications has been studied by academia for several years [22, 28], very few precoding techniques have been reported by the commercial companies designing HTS systems. Rather than considering the ag- gressive UFR scheme enabled by advanced precoding and user scheduling techniques, the 4CFR multibeam scheme is still the main consideration of the upcoming commercial HTS systems, such as ViaSat-3$^{@}$ [64, 65], Eutelsat Konnect VHTS$^{@}$ [66], SpaceX, and Tele- sat [48]. To our best knowledge, UFR in combination with precoding and user scheduling has not been employed in the state-of-the-art commercial HTS systems [48, 67], although research has shown its great potential.

Nevertheless, we can find some ongoing projects exploring the implementation of pre- coding in real GEO satellite systems, such as [68] carried out by researchers at SnT, University of Luxembourg, and supported by European Space Agency (ESA). In 2019, researchers with SnT investigated the feasibility of implementing multibeam precoding in a UHTS SatCom system, using software-defined radio platforms [69–71]. The hardware- based experimental results show that it is feasible to develop precoding under an UFR scheme in industrial applications. Therefore, we believe that there are enormous research value and potential markets to develop this technology in the space industry.

2.7 Key Findings and Future Research

- The previous sections show that the aggressive universal frequency reuse (UFR) scheme can potentially increase the degree of freedom by a factor of 2 to 4. Com- pared with the conventional four colour frequency reuse (4CFR), UFR can improve the total throughput of the HTS by about 20% to 40%, even when only quite basic linear precoding methods are used. Since UFR will introduce multibeam interfer- ence, multibeam precoding is essential to mitigate the interference in HTS systems.

- When user scheduling is jointly used together with multibeam precoding, the total throughput can be further improved by about 10% to 60% (as shown in Table 4), compared with the methods without user scheduling.

- The performance of precoding and user scheduling in HTS can be considerably degraded under realistic conditions, such as rain attenuation, outdated CSI, and the limited feeder link bandwidth. Therefore, robust multibeam precoding and user scheduling design remains a critical issue for future HTS.

- Regarding the limited feeder link bandwidth, implementing multiple GWs and hybrid on-board/on-ground precoding can be effective solutions. The architecture of GW networks and the connectivity and cooperation between GWs has a great impact on the satellite throughput and the QoS in ISTNs. However, the study of implementing multiple GWs is insufficient, especially when taking the real-word issues into account. Besides, how the architecture of GWs can influence the performance of precoding is still unclear.

- The adaptive multibeam processing techniques can further improve the performance of HTS in terms of providing satisfactory QoS for users with efficient utilization of limited resources. The existing research on adaptive precoding considering traffic demands under UFR scheme is insufficient, especially when multiple users served by one beam. Besides, unbalanced traffic demands for each beam can possibly cause insufficient and excessive supply, and user scheduling can balance the capac- ity requirements for each beam. However, there is currently no relevant study on multigroup multicast scenario implementing joint precoding and user scheduling matching traffic demands, which can be further combined with the adaptive spot- beam generation in terms of flexible beamwidth, pointing direction, onboard power allocation and dynamic coverage.

- The adaptive multibeam processing techniques can further improve the performance of HTS. With the adaptive design of precoding coefficients by jointly considering the resource allocation and interference mitigation, these techniques are promising in terms of providing satisfactory QoS for users with efficient utilization of limited resources.

- There is a clear industry gap between the state-of-the-art HTS systems using the 4CFR scheme, and the HTS systems studied by academia exploiting UFR scheme together with multibeam precoding and user scheduling techniques. The latter can be a great asset for the future HTS systems in terms of further significantly improving the throughput, mitigating the interference and providing various needs to users. To bridge the industry gap, practical and pragmatic multibeam processing which considers all real system constraints, such as per antenna power constraint, rain attenuation, delayed channel station information, uneven multibeam loads, etc., need to be investigated and tested jointly with industry partners.

- There is a novel and innovation development on dealing with high-Doppler frequency shifts for satellite communications, particularly for LEO satellites systems. How to compensate for high-Doppler frequency shifts is an important issue for practical systems. In this aspect, multibeam precoding schemes designed in the delay-Doppler domain based on principles of new Orthogonal Time Frequency and Space (OTFS) scheme is very promising, in order to provide Doppler-resilient performance in a very dynamic changing environment. It is compatible with existing 5G or OFDM based signals and also has low complexity to scale the capacity and throughput for multibeam systems.

3 Adaptive Spot Beams and User Scheduling to Meet Traffic Demands

For spot-beam HTS systems, numerous adaptive algorithms have been proposed to adapt satellite resources in response to user demands. A large number of these algorithms do not include precoding which is an established technique for reducing inter-beam interference and widely proposed for terresterial networks. It is reasonable to expect that precoding will also be used in future HTS systems.

We now proceed to present material reviewed in the following sections. Section 3.1 looks at adaptive beam hopping techniques. Beam hopping is attractive because it pro- vides means to share bandwidth between various cells within the coverage area. Conven- tional fixed-beam systems do not allow such sharing of bandwidth and rely only on beam size to determine resource allocation. Most of the papers we discuss consider joint beam hopping and precoding algorithms.

In Section 3.2 and 3.3 we consider interference-aware scheduling for the forward and return link, respectively. The approach in each case is to schedule so as to either avoid interference or control it. Precoding for the downlink is not considered.

Section 3.4 looks at multiple gateways which can be used to improve the reliability of the feeder link and to increase the system throughput. Traffic losses can occur due to rain attenuation of the feeder link or due to a gateway failure. We review adaptive algorithms designed to switch between the gateways to provide reliability in the face of rain attenuation and gateway failure. We also consider the question of how to match user traffic to the gateway feeder links.

Sections 3.5 and 3.6 present papers which are concerned with adaptive allocation of power and bandwidth. This can be done according to various criteria. Crucial amongst these are those which use some form of traffic estimate to formulate the adaptation ob- jective, e.g., minimize the squared error between capacity and traffic demands.

3.1 Beam Hopping Techniques

In current broadband multi-beam HTS systems, all satellite beams are constantly illuminated regardless of the traffic demand. As a result, there is inefficient utilization of available resources since the demands of service or capacity may vary over time and across different geographic areas.

Beam hopping enables flexible allocation of resources, such as power and bandwidth over the service coverage area, allowing traffic variations to be addressed. Beyond this, interference can be managed through the choice of beam hopping pattern and cells to be illuminated. Nevertheless, when the high throughput UFR pattern is employed among beams, the performance of beam hopping will be degraded, due to inter-beam interference. Hence, combining precoding and beam hopping is an attractive solution to this problem.

[72] considers beam hopping over “snap-shots”. Here the objective is to directly match resources to user demands over a beam hopping cycle. This is a discrete problem of time slot allocation to snap-shots as part of a mixed integer linear program. They later consider relaxations to a linear program yielding an upper bound. To solve this problem they propose a neural net classifier to identify snapshots to be used in real time reducing the problem to a linear program which determines the time allocated to each snapshot in the cycle.

A joint precoding and beam hopping approach was developed in [73], employing linear zero-forcing (ZF) precoder with on-board per feed power constraints. The simulation results show that about 65% throughput gain is obtained by the joint precoding and beam hopping method, compared with the 4CFR scheme. Optimisation is with respect to functions of SINR and could incorporate traffic demands as weights for example.

Recently, an interesting concept of cluster hopping was proposed in [74]. By using cluster hopping, the set of clusters, i.e., a set of adjacent beams, are predetermined and illuminated simultaneously at each hopping event, along with the illumination duration. The use of clusters reduces the problem of beam allocation to be of smaller size. At the same time precoding can be carried out at each cluster to further increase throughput. Fig. 13 shows the performances of different resource allocation schemes in terms of cluster demand and offered capacity. It can be seen that compared with the 4CFR and FFR schemes that offer almost unchanged capacity, the proposed cluster hopping method exhibits efficient rate matching capability.

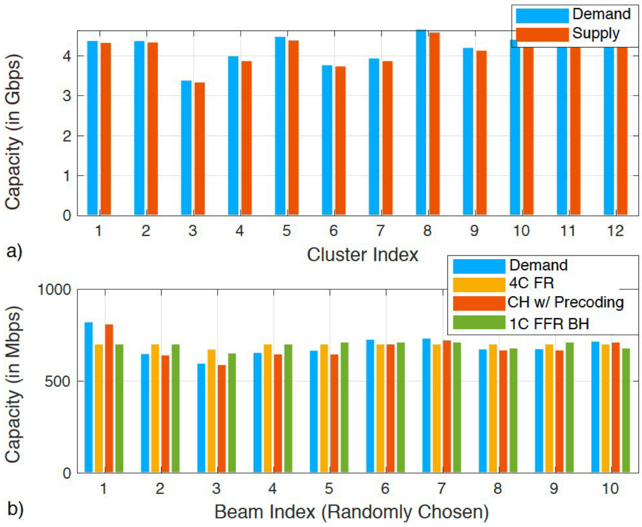


Figure 13: a) Demand capacity vs. offered capacity at cluster level, and b) demand vs. offered capacity at beam level (for a set of randomly selected users) [74].

[74] provides parameters for beam hopping of a roughly 1 ms. dwell time over 256 time slots to give $1/4\ s$ = 1 ms 256 hopping cycle. Note that beam hopping leads to additional delay as well as possible delay jitter. Some protocols e.g. 5G protocols may be adversely affected as a result. Moreover there is increased latency on the same order as the propagation time from the gateway(s) to users in a GEO system.

We now add a few additional remarks on beam hopping. Obviously, the choice of beam hopping pattern is crucial in assigning the correct amount of bandwidth so as to match the underlying traffic demand. None of these papers consider how these demands are to be estimated. However they do suppose that this is done on a per cell basis. Also, they do not state over what timescale adaptation should take place e.g. every second, every minute or every hour, each coming with its traffic estimate. Traffic is considered only as an aggregate over the cell.

It should be further noted that beam hopping techniques are confined to hopping over a given fixed set of beams. This is consistent with conventional feed horn/reflector antenna systems in current use. However, future designs are unlikely to use such fixed- beam designs. Instead, it is expected that antenna beams will be based on phased arrays. Such antenna arrays have considerably more flexibility than simply prescribing beam hopping patterns with some additional power and bandwidth assignment. Phased arrays allow the beam locations and beam sizes to be adapted as needed, for example to manage interference and also match traffic more effectively.

3.2 Interference-Aware User Scheduling in the Forward Link

In [75], an interference-aware scheduling algorithm was presented for the forward link. In brief, this scheme aims to have higher reuse by careful progressive selection of users in different beams. The scheme also has fairness criteria built in to ensure that there is no user starvation.

In more detail under this algorithm each user feeds back partial channel state information, i.e. power received by the given user from each beam. The idea is then to progressively schedule users over a series of scheduling steps such that at each time step, interfering beams are deactivated before a new desired beam is selected.

Simulation results were given for a 302 beam layout for Western Europe. Mean rates for a population of 17986 users are obtained - CDFs for users and beams, see Figure 2 in [75]. The benchmark scheme is a scheme using frequency reuse 4 and round robin scheduling. The results are for per user spectrum and per beam spectrum efficiencies. Results yield around 70 Mbits/s on the Forward link per user for IAS and around 55 Mbits/s per user for the reuse 4 round robin benchmark.

As with beam hopping discussed earlier, this algorithm is for application with con- ventional fixed-beam systems. The question of how to schedule users in conjunction with precoding is not considered in [75]. However, this consideration is addressed in Section 2. It remains to be seen whether and how such a scheme might be used in conjunction with phased array antenna systems. Furthermore, there are various modifications that can be made with the broad scheme described in [75], which might also be considered.

3.3 Interference-Aware User Scheduling in the Return Link

In the forward link, the interference at a given user mainly depends on the user's own location with respect to the interfering beam patterns, where as in the return link, the interference received by a given user signal depends on the position of interfering users in the interfering beams as shown in Fig. 14.

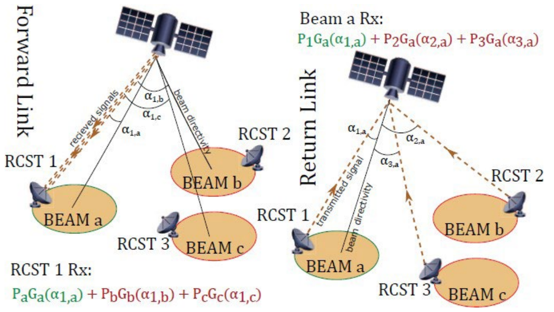


Figure 14: The difference between interference in forward link and return link [76].

[77] explored the achievable rates of successive inference cancellation (SIC) for the return link of multiuser spot-beam satellite systems. Only one user per beam was assumed. The resulting multiuser return link was considered to be equivalent to a multiuser MIMO

uplink, for which SIC was applied. The performance of SIC was compared with multiuser access with neither interference cancellation nor receiver combining, and multiuser access with maximal-ratio combining at the satellite.

The results in [77] are theoretical based on the maximum sum rate for MIMO multi access channel. However, the performance depends on the selection and ordering of users making brute force selection unworkable. To address this problem, the authors examine various heuristic schemes for determining user selection and ordering.

It is clear from this paper and others, that the schedule of users onto the return link determines the capacities of the corresponding multiple access channels, and therefore the achievable user throughputs. Scheduling therefore is an inevitable consideration in optimizing user throughputs.

[78] also applied the principles of multiuser MIMO uplink for the return link of spot- beam system and SIC was used to decode user signals. A time-division multiple access (TDMA) scheme was considered for each beam. With M users per beam and B beams, the optimal user scheduling scheme has to search through $(M!)^{B-1}$ schedules, which is computationally infeasible as M and B increases. Therefore, a heuristic user scheduling scheme which used multi-partite graph matching was proposed to maximize the minimum user rate. Here the UTs are divided into groups each of which is represented by a color (this is a virtual color). These are groups which would have high SINR if scheduled together. They can be transmitted on one of M time slots. Users in the same beam must transmit in distinct timeslots but otherwise the schedule is random. One such schedule is obtained for each color i.e. group of users. Having obtained these partial schedules, the complete schedule is obtained by merging them. This is done through a matching algorithm. Say there are only two colors. Then each time slot defines a subset of users and this may be paired with another subset. The weight between these two pairs is the minimum of the user rates, across the subsets. The minimum rate is maximised by solving the corresponding matching problem. This is solved using a greedy algorithm. If there are C colors (groups) this process is repeated $C \quad 1$ times.

A heuristic interference coordination scheme for the spot-beam multi-frequency TDMA (MF-TDMA) return link was proposed in [79]. The objective of the proposed scheme was to maximize the minimum carrier-to-interference ratio (CIR). The performance of the proposed algorithm was evaluated in a 302 spot-beam layout across Europe with 200 users uniformly distributed in each beam. The complementary cumulative distribution function (CCDF) results of CIR show that the proposed interference coordination scheme outperforms the static fractional frequency reuse scheme.

Interference-aware frame optimization algorithms for the DVB-RCS2 spot-beam MF- TDMA return link was proposed in [76]. Each MF-TDMA frame consists of multiple time and frequency channels, which are called as Bandwidth Time Units (BTUs). The DVB- RCS2 standard allows applying modulation and coding (ModCod) schemes independently for each BTU. The DVB-RCS2 standard also defines a set of ModCod schemes set with an SINR threshold for each modulation scheme (Table II). The aim of the proposed frame optimization algorithm was to find the user schedule and ModCod schemes to be used for each BTU in each beam in order to maximize the total system throughput. The authors also proposed sub-optimal time-wise decomposition and subcarrier-wise decomposition algorithms as well, since the solution space of the original Global optimization problem was prohibitively high. The results show that even the proposed decompostion algorithms, which are sub-optimal, achieves about 50% gain in total system throughput, compared to the four-color frequency reuse scheme.

In [80], two interference-aware user scheduling algorithms were proposed for the DVB- RCS2 spot-beam MF-TDMA return link. Unlike [76] ModCod optimization was not considered and all the users were assumed to be using the same ModCod scheme. In the Greedy scheduling algorithm, in a given BTU, the user with the minimum interference is scheduled, without considering the interference added by the new user to the currently active users. In the Fair scheduling algorithm, the user which adds minimum interference to the existing users is scheduled to the given BTU. Interestingly, the results show that the both algorithms resulting in similar sum throughput values. However, the computational complexity of Fair scheduling is slightly higher than Greedy scheduling. The authors of this work claimed that their proposed algorithms completed within 25 ms while the algorithms in [76] took 50 s.

[81] also proposed a user scheduling algorithm for the spot-beam MF-TDMA return link. They considered both co-channel interference and adjacent channel interference (ACI), where ACI was modeled as a scalar parameter β. In their model, each user in each beam requests a BTU (termed as Frequency-Time Quanta (FTQ) in the paper) with a given probability, which is the same for all the users. The authors sought to find the user schedule which maximizes the sum throughput, both with and without successive interfer- ence cancellation. The computational complexity of exhaustive search over all the users to find the optimal schedule increases exponentially with the number of users and the num- ber of beams. The authors proposed a scheduling algorithm based on genetic algorithms, which reduced the number of iterations significantly. With SIC, the sum throughput did not vary with the ACI parameter β. However, without SIC, the ACI parameter had a significant impact on sum throughput. Furthermore, the proposed algorithm converged to the sum-throughput achieved by exhaustive search using a significantly lower iterations than exhaustive search.

In [14], the performance of four-color frequency reuse, four-color fractional frequency reuse (FFR), four-color partial frequency reuse (PFR), four-color soft frequency reuse (SFR) and two-color frequency reuse were compared for the return link. Four per- formance metrics were considered, the maximum throughput, throughput achieved by proportionally-fair resource allocation, throughput with coordinated modulation and cod- ing scheme selection, and throughput without any interference coordination. For SFR and PFR, the throughput of coordinated schemes increases as the power ratio between the beam cell users and the beam edge users increases. The two-color frequency reuse showed the best and worst total and per-user throughputs for the coordinated and uncoordinated schemes, respectively.

The main idea of the Two-color interference coordination scheme in [11] is to use a higher frequency reuse than in a conventional four color scheme. Making the reuse more aggressive increases the available bandwidth but at the expense of higher interference.

Here proportional fair (PF) scheduling is proposed and user requests for resources are received. (Users are constrained to transmit on only one carrier at a time according to the standard.) Once the requests are in assignments are made to carriers and times slots and then the schedule is run. Power control is slow it is argued and therefore operation is with fixed power.

Under a benchmark scheme interference is estimated using a tunable scale parameter η., and overall interference power, see (5). Depending on the choice of η we will lose many data blocks (underestimate interference) or transmit at too low a rate (overestimate interference).

The scheme proposed makes this initial assignment step. Having obtained the schedule the interference can be computed exactly and a new choice for MODCON be determined for each user and timeslot. By taking this step the authors obtain a 50 % increase in throughput, see table II, page 999 and Figure 6 page 1001.

A more complex problem is to make joint scheduling and MODCON selection together - not in the two step process above which is clearly subooptimal. The authors simulate this approach which yields throughput gains on the order of 80 %.

Again, none of the papers above consider scheduling users according to traffic demands. The central problem identified in these papers is to obtain the best user schedule. This is a problem of high complexity, and therefore heuristics are applied. There is no discussion of performance bounds for these schemes for the return link. Also, once again the papers focus on conventional fixed-beam systems. The additional flexibility, etc of phased arrays is not taken into account.

3.4 Gateway Diversity and Multiplexing Schemes

Clearly, the feeder link is vital to the end-to-end communication, and gateway diversity schemes with multiple gateways can be employed to improve the availability of the feeder link. Typically an availability higher than 99.9% should be targeted for the feeder link. Furthermore, gateway multiplexing can be used to increase the feeder link bandwidth to serve more user terminals.

In the system with multiple feeder gateways considered in [82], in each time slot, a given feeder link is only routed to one user beam in the forward link. The objective of the adaptive gateway switching algorithm is to match the traffic offered by the gateways to the traffic requested by the users in each beam. Three different objective functions are considered for the switching algorithm, rate matching, load balancing and fairness, which are given in eqs. (3), (4) and (5) in the paper, respectively. It was shown that compared to the case with a fixed frequency division between feeder links, the capacity losses occur due to deep fades in feeder links can be significantly reduced by the proposed three time-switching algorithms.

In [83], the improvement of gateway availability achieved by adding P redundant gateways into the N-gateway architecture $(P \ll N)$ in [82] was evaluated. It was shown that adding one extra gateway per four, six, or eight active gateways, the feeder link will be available for 99.98 or 99.96% of the time, respectively. The gateway availability increases to 99.999% by with two redundant gateways. Furthermore, the proportion of redundant gateways to be added to achieve a given availability value reduces as the size of network increases.

Fig. 15 shows the architecture of the gateway diversity scheme proposed in [84] for spot- beam satellites, where each gateway is connected to a software-defined networking (SDN) switch and all the SDN switches are controlled by a SDN controller using the OpenFlow protocol [85]. C traffic priority classes which require different guaranteed throughputs were considered. The authors proposed a traffic handover algorithm to be implemented at the SDN controller. The proposed algorithm consists of two main components. The Traffic flow association component assigns incoming traffic from each traffic class to avail- able gateways to satisfy the guaranteed throughputs for each traffic class. In the Traffic flow reallocation component, if the SNR of a feeder link decreases, the SDN controller reruns the Traffic flow association component with the remaining available gateways. The handling process of existing queues in the gateway with outage is also described in the paper.

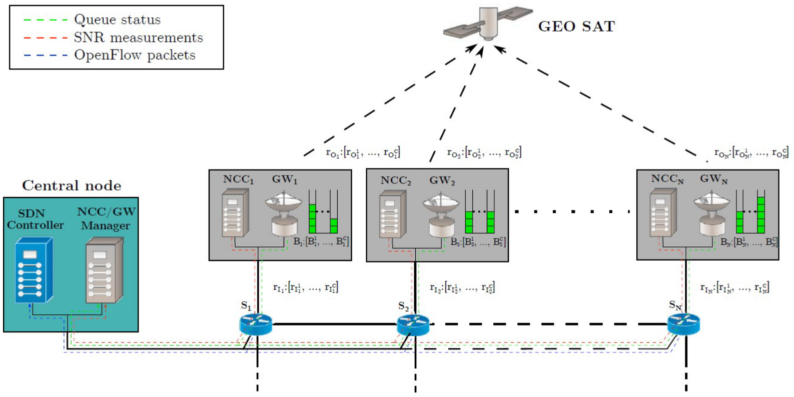

Queue status
SNR measurements
OpenFlow packets
GEO SAT
r_{O_1}:[$r_{O_1}^1$, ..., $r_{O_1}^c$]
r_{O_2}:[$r_{O_2}^1$, ..., $r_{O_2}^c$]
r_{O_N}:[$r_{O_N}^1$, ..., $r_{O_N}^c$]
Central node
SDN Controller
NCC/GW Manager
NCC$_1$
GW$_1$
NCC$_2$
GW$_2$
NCC$_N$
GW$_N$
B_1:[B_1^1, ..., B_1^c]
B_2:[B_2^1, ..., B_2^c]
B_N:[B_N^1, ..., B_N^c]
r_{I_1}:[$r_{I_1}^1$, ..., $r_{I_1}^c$]
r_{I_2}:[$r_{I_2}^1$, ..., $r_{I_2}^c$]
r_{I_N}:[$r_{I_N}^1$, ..., $r_{I_N}^c$]
S$_1$
S$_2$
S$_N$

Figure 15: The architecture of the gateway diversity scheme proposed in [84].

A combination of gateway diversity and multiplexing was considered in [21]. The target was to achieve 1 Tbps system capacity for the continental Europe while maintaining an availability of 99.9% for the feeder link. Feeder links were operated in Q/V band with universal frequency reuse between feeder links. 1 Tbps capacity could be achieved using 190 user beams, which were shared between 19 gateways.

3.5 Adaptive Power and Bandwidth Allocation

A traditional approach to dynamic resource allocation in terrestrial cellular systems is joint power and carrier allocation. Joint power and carrier allocation for the downlink of a multi-spotbeam satellite system is considered in [8]. This is formulated in terms of SINR requests per beam (not user). This involves a spectral mask for each beam over the various carriers indicating which carriers are used and how much power is assigned.

Rates are defined via the SINR's for the k^{th} beam as the sum of the respective Shannon rates. An iterative scheme is used to assign carriers until either all SINR requests are met or the power constraint is exceeded.

An algorithm was proposed in [86] for adaptive allocation of bandwidth and power between the forward downlink (downlink between the satellite and user terminals) and return downlink (downlink between the satellite and the gateway). In this system, the satellite transponder decodes and forwards the signal from the gateway to the user ter- minal and vice versa. The objective of the power and bandwidth allocation algorithm in [86] to maximize the capacity of the return downlink subject to the constraint that guaranteed transmission rate requirements of the forward downlink are satisfied.

In [87], packet scheduling techniques to be implemented at the gateway were developed to accommodate carrier aggregation in the forward link of DVB-S2X systems for a single user. The carrier aggregation was transparent to the layers higher than the Link layer. With carrier aggregation, the packet scheduler distributes the incoming packets from the higher layers between the two carriers aggregated based on their channel conditions (a load balancing scheduler). Furthermore, a packet allocation technique was proposed to ensure that the traffic merging block at the user terminal is simply a first-in first-out buffer.

3.6 Traffic Matching

Analysis for network dimensioning is presented in [88], see e.g. chapter 12. This includes capacity calculations and traffic breakdowns together with case studies. Figure 12.4 in [88] shows the distribution of traffic across a set of spot beams for the SECOMS satellite system. For the downlink it can be seen that a small number of beams carry around 150 Mbits/s whereas the majority of beams carry around 50 Mbits/s or less.

Such results are usually aimed at busy hour periods and therefore to peak traffic requirements. This is a reasonable approach for a system with fixed beams and reuse patterns. However it neglects the performance gains which can be made by throughput matching. The considerable variation in traffic over time and at different locations suggest that even an hour by hour adaptation of bandwidth and beam "profiles" would lead to significant performance gains.

For this book and the papers cited below there is little or no discussion of delay which is an important consideration for new services including those for 5G. The question of how to manage mixed QoS requirements is also not addressed for the most part. Of course the need to address delay and also other traffic constraints can only reduce the throughput which could otherwise be supported.

The papers below consider the question of how to translate beam by beam throughput requirements to resource allocation e.g. power and bandwidth in fixed beam systems. In all these papers the question of how to determine the required throughputs is not addressed nor is there much discussion of the timescale on which resources are being adapted.

The papers [89] and [90] both use the same objective function for matching beam capacities to target throughputs. That is to choose the rates C_i so as to best match a set of target throughputs T_i according to a sum-of-squares criteria:

$$\min \sum_{i} (T_i - C_i)^2 \tag{3}$$

Both papers address this problem through the method of Lagrange multipliers and as- sume that the traffic vector $\mathbf{T}$ is given. Both papers neglect interference between beams. An algorithm is presented in [89] to determine the optimal joint power and bandwidth allocation. [90] uses similar methods to determine the best K beams to use and optimal bandwidth allocation. Power is fixed. The paper [57] uses the same methods and objective function to determine the optimal power allocation and presents an iterative algorithm.

Interbeam interference is included in [91] using the same objective (3) and is addressed via an interbeam coefficient matrix. Delay is considered by transforming delay into a throughput requirement providing a rough allowance. How to tradeoff between the QoS requirements is not considered. The problem is to again minimise the objective given in (3) over power and bandwidth. This is not a convex optimisation problem as the power terms additionaly appear in the interference, unlike the previous papers. However as previously the solution is obtained iteratively using duality theory.

A mixed fair sharing, capacity mismatch objective is formulated in [92], the latter again can be taken as in (3). The paper considers 3 different payloads with different degrees of flexibility: in the first both power and bandwidth can be assigned; in the second only bandwidth; and in the third only power. The optimisation problem is presented on page 268 of [92]. This is a non-convex optimisation problem which is solved using Simulated Annealing. Interference is not neglected.

A scanning (beam hopping) satellite system, including the case where there are fewer beams than cells, is considered in [93]. As above they use objectives such as given in (3).

The objectives are optimised for power and bandwith as in the earlier papers. However [93] also considers delay in terms of that acquired through packet transmission error. They also investigate the performance impact of using a limited number of beams.

Optimal power allocation based on traffic demands for a multi-beam satellite systems is considered in [94]. This has been formulated as an optimization problem with the twin objectives of traffic matching and fairness maximization. Fairness is addressed using an index and is aimed at addressing the asymmetry of the power allocation. Optimisation is performed by adopting a multi-agent construction using reinforcement leaning. There is one agent associated with each beam. At each stage there are state transitions which are done simultaneously for all the agents. The state itself consists of a history, the channel conditions and traffic demand. The approach is akin to simulated annealing in that there is a temperature and action probabilities are determined via a Q function. The increments of this Q function are obtained using a rate reward and a conjecture function, see equations (11) and (12) in [94]. These are iterated with a learning rate constant as well as being discounted.

The paper [58] formulates the twin objectives of matching the target throughputs whilst at the same time minimising the power. This is the least power needed to meet a given throughput vector. The problem is addressed using a two stage heurisitic algo- rithm. The first problem is a Multi Objective Optimisation Problem (MOP) aiming at determining the beam power levels P_b that satisfy the respective traffic demand and at the same time minimize DC power consumption. The second stage of the optimization aims at enhancing the solutions obtained in the first stage with regard to power utilization. It also aims at determining the relevant Pareto front related to the MOP.

3.7 Key Findings and Future Research

In addressing the problem of adaptive spot beams, we are often confronted with large scale multi-dimensional optimisation problems owing to the large number of possible an- tenna configurations together with the possible power and bandwidth allocations. These problems are prohibitively complex so obtaining the optimum in real time is impractical. Since this is the case, efficient solutions must be developed using heuristic approaches and other techniques. In this regard, theoretical results such as performance bounds and solutions for idealized cases will be needed.

Below a number of issues for adaptive beam systems are presented. These represent gaps in the literature where there are few or no results available, and yet appear crucial for future satellite systems.

- Future antenna systems will likely be based on phased arrays and not on fixed beam systems, which the majority of the literature addresses. Phased array systems have considerably more flexibility allowing both beam size and beam location to be adapted as needed. Moving beams are a further consideration for some applications e.g., cruise ships, planes and trains.

- Future scheduling schemes on the return link should be traffic based. It is therefore crucial to understand performance under a broad set of scenarios and furthermore not depend solely on aggregate metrics such as sum throughput, as considered so far in the literature. Thus, user demands must be incorporated into schedules alongside other criteria.

- In satellite systems with adaptive spot beams, the problem of matching traffic to gateways is more complex than heretofore considered. In fixed spot-beam systems, the problem of association of gateways to the beams to fulfill traffic demands is not a complex problem since it is a matter of allocating a predefined set of feeder links to a predefined set of beams based on traffic demand of beams and feeder link channel gains. This is not the case for adaptive spot beams. Novel efficient algorithms which do not incur large delays are required to obtain optimal beam configurations and gateway associations for adaptive spot beam systems.

- The issue of matching capacity to traffic has only been addressed on a per-cell aggregate level. Moreover, there is little work on how traffic should be estimated. Obviously, this is a crucial consideration. Also neglected is the time scale on which resources should be adapted to the underlying traffic demands. Adaptation on hour-by-hour or even minute-by-minute basis is achievable in practice. But the algorithms to do this and the mechanisms to coordinate user information and resource allocation are yet to be identified. As a last point, working at the level of per-cell traffic aggregation may not be efficient in meeting user quality of service. Allocation of bandwidth, etc, should take into account user traffic variations, and not just mean rates.

4 Adaptive Spot Beams with LEO Satellites

In this section, we will review the LEO satellite network architecture underpinning adap- tive spot beams technology and associated design challenges. We will start with a brief overview of new generation LEO satellite systems.

4.1 New Generation LEO Satellite Systems

The new generation LEO satellite systems featuring adaptive spot beams are currently on the cusp of massive proliferation and deployment in space, with a steady increase in the number of proposals filed by various space, telecommunications and information technology companies across the globe in recent years, e.g., more than 11 proposals filed to US Federal Communications Commission (FCC) from 2014 to 2018 [48]. It is forecast that there will be 50000 satellites orbiting the Earth in 10 years if the current satellite Internet proposals become reality [95]. Among the recent fillings, SpaceX proposes to have 4425 LEO satellites using Ku-Ka bands with target latency from 25 to 35 milliseconds[1], while Telesat aims to deploy a 117-satellite Ka-band constellation at altitudes of 1000- 1200 kilometers with target latency from 30 to 50 milliseconds [96]. Amazon has also recently announced its plans to invest USD 10B to deploy 3236 LEO satellites under the project named *Kuiper*.[2] We will refer to satellite constellations containing more than 100 LEO satellites as *big-LEO satellite constellations*. Two big-LEO constellations proposed by SpaceX and Telesat are illustrated in Fig. 16.

These new generation LEO satellite systems will require significant investment in both *space* and *ground* segments. The space segment will consist of hundreds of LEO satellites orbiting the Earth and optical communications links connecting satellites in space to form a dynamically changing mobile network topology. The ground segment will consist of gateways (providing an interface to Internet), ground control center (for computing intensive network control and planning functions such as routing) and consumer-premise equipment (with electronically steerable antennas to track the satellites). The gateways are connected to each other to form a virtual private network. Compared to GEO HTS communication systems, they have the advantage of providing smaller latency, better channel attenuation and higher elevation angles at high latitudes as well as having lower production and launch costs [48, 95, 96].

The new generation of LEO satellite systems presents great potential for broadband connectivity as well as provisioning remote IoT services. As reported in a recent study [48], SpaceX system is capable of delivering maximum throughput of 23.7 [Tbps] (with more than 100 ground stations), while the maximum throughput is around 2.66 [Tbps] for Tele- sat with 42 ground stations. To put the potential presented by these LEO satellite systems into perspective, it is also worthwhile to note that the per-user rates delivered by them are expected to be higher than pre-5G terrestrial wireless rates [95]. Beyond broadband services, Fleet (https://www.fleet.space/about) and Myriota (https://myriota.com/) are two example Australian startup companies, in rapidly changing and life-transforming IoT sector, utilising LEO satellites to provide connectivity for IoT devices in remote locations lacking terrestrial network coverage. The main technological advances unlocking the po- tential of LEO satellites and leading to proliferation of big-LEO constellation proposals

[1] SpaceX has also another approval from FCC to have additional 7518 satellites located at altitudes just below 350 kilometers and using V-band spectrum.

[2] The market anticipates that Facebook and Apple will also join in the race to deploy big-LEO constellations in space [97, 98].

Figure 16: Planned big-LEO satellite constellations from SpaceX (left-hand side figure for 4425 LEO satellites) and Telesat (right-hand side figure for 117 LEO satellites). SpaceX proposes to have 5 different set of orbital planes at different inclination angles, whereas Telesat plans to deploy polar (red) and inclined (blue) orbits at inclination angles 99.5° and 37.4°, respectively. Modified from [48].

and startup companies, as opposed to previous failed attempts in 1990s including Global- star, Iridium, Odyssey and Teledesic, are digital payloads supporting software-define radio technology, multi-beam antennas, advanced modulation and dynamic resource allocation techniques.

Next, we will focus on the LEO network architectures supporting the use of more advanced and dynamic antenna techniques and then discuss the challenges pertaining to radio resource management and network control due to peculiar channel and operating conditions for LEO satellite uplink and downlink channels. We will call a satellite LEO satellite if it orbits the Earth at an altitude ranging from 500 to 2000 kilometers [96], while noting that SpaceX's latest proposal approved by FCC aims to deploy satellites at altitudes just below 350 kilometers, which they call very low-Earth-orbit (VLEO) satellites.

4.2 Network Architecture and Design Challenges

Three notable big-LEO satellite constellations approved by FCC are SpaceX's Ku-Ka band system (having 4425 LEO satellites), Telesat's Ka band system (having 117 LEO satellites) and OneWeb's Ku-Ka band system (having 720 LEO satellites). Among these three, OneWeb has recently ceased its operations due to COVID-19 pandemic, and therefore it will not be a focus of our discussion.

The Telesat's proposed constellation consists two sets of orbits. Six polar orbits, represented by red in Fig. 16, will be at 1000 kilometers with an inclination angle 99.5° and consist of at least 12 satellites per orbital plane. The second set is the inclined orbits (more than five), represented by blue in Fig. 16, located at 1200 kilometers with inclination 37.4°, and containing at least 10 satellites per orbit. The logic behind the proposed design is to provide general global coverage with polar orbits, while simultaneously covering mostly populated regions on the Earth with inclined orbits. SpaceX plans to have a more complicated big-LEO constellation than that of Telesat to deploy 4425 satellites in space, which is illustrated in Fig. 16. In this planned constellation, there will be 5 sets of orbits at altitudes and inclination angles ranging from 1110 to 1325 kilometers and 53° to 81°, respectively. We note that the big-LEO constellations by SpaceX and Telestar illustrated in Fig. 16 are also known under the name of Walker-Delta constellations in the literature [96]. An important feature common to both systems is the divergence from bent-pipe payloads to digital ones allowing to implement physical (e.g., phased array beam steering, modulation, demodulation, on-off interference avoidance) and some potential network layer capabilities (e.g., congestion control) on-board in LEO satellites. This is an important paradigm shift from their predecessors proposed in 1990s. In addition, both systems will feature inter-satellite links (ISL) in order to provide continuous connectivity even when a user and a gateway are not simultaneously within the line-of-sight of an LEO satellite.

Three main technical challenges in the way to make these systems operational are interference coordination between between GSO and NGSO orbits, dynamic radio resource management (DRRM) to maximize throughput and management of frequent handovers due to high satellite mobility (i.e., usual LEO satellite speeds at 5 to 10 km/s will trigger a handover in less than a minute for a typical LEO spot-beam having radius 450 kilometers) [48, 96, 99]. With continuing advances in antenna technology, one can further envisage to have beam footprints with radius smaller than 450 kilometers to increase frequency re- use and improve system throughput but this will place an added burden on the system design with much more frequent handovers, e.g., a handover approximately in each 6 seconds with 45 kilometer beam radius. Electronically steerable antenna (ESA) elements at user premises, satellites and gateways can be used to increase dwell time within the coverage area of a spot-beam, and thereby to alleviate the frequent handover problem. An elaborate beam-hopping design to handover the user among the beams of a satellite will also have the similar effect by keeping a user connected to the same satellite for a longer time duration. In the case of DRRM, it is imperative to have efficient implementations for scheduling of frames, frequencies and bandwidth, as well as allocation of transmitted powers among the beams and dynamic steering of beam directions [92, 100]. The problem of effective distribution of network algorithms between ground and space segments is also important to solve as some of these algorithms will need to run on-board to adopt to the rapid changes in the propagation environment and some will need to run at a ground control center (i.e., network management center) having a general overview of the big-LEO constellations, gateways and user demands.

The spectrum usage patterns show some notable differences for SpaceX's and Telesat's systems. Telesat's system will only use the Ka-band for both user and gateway communi- cations with downlink and uplink communications taking place at the lower Ka-band from 17.8 to 20.2 GHz and upper Ka-band from 27.5 to 30.0 GHz, respectively. On the other hand, SpaceX's system treats user and gateway communications differently. It allocates Ku-band 10.7-12.7 GHz for user downlink and 14-14.5 GHz for user uplink communica- tions. For gateway communications, it uses Ka-band 17.8-19.3 GHz for gateway downlink and 27.5-30 GHz for gateway uplink.

In both cases, the ground segment will consist of gateways and consumer-premise equipment, having ESA elements. In case of Telesat, they expect to have several gate- ways geographically distributed over the Earth, whereas SpaceX plans to have a very large number of gateways. An important bottleneck here is the cost of consumer-premise equipment. Traditional parabolic-dish antennas are not suited to track LEO satellites due to non-geostationary nature of these rapidly moving satellites. What is needed is ESAs shifting beams without any physical movement to track LEO satellites. The cost of hav- ing an electronically steering antenna is, however, estimated to be in the order of several thousand dollars [95]. These costs are not bearable to tap into the consumer market and must be lowered an order of magnitude without sacrificing from high data rates, reliable beam steering and smooth antenna handover. Table 5 provides a comparison between Telesat's and SpaceX's big-LEO constellations.

Satellite link characteristics, constellation orbital information, ground segment config- uration and user demand maps are important system-level parameters to determine the peak and average data rates that can be delivered by these big-LEO constellations. To this end, a statistical model is developed in [48] to estimate the system coverage and through- put (i.e., sellable capacity), with ground segment optimization by means of a genetic algorithm to determine the locations of gateways. Among many other results, it is shown that both systems achieve more than 90% direct coverage (defined as the percentage of points covered by a satellite that can simultaneously talk with a gateway) by using at least 40 ground stations. There are two important remarks about these estimated coverage val- ues. The first one is that they are based on the assumption of homogeneous spot-beam coverage model. By adopting a hierarchical hybrid spot-beam coverage model, having a mix of concentrated and wide spot-beams, the coverage efficiency of LEO satellite sys- tems can be improved further in a way more conducive to network management [96, 101]. Secondly, these systems are equipped with ISLs to route data from satellites out of the coverage area of a gateway to those that can simultaneously communicate with a gateway, or vice versa. Hence, even if a location is not within the direct line-of-sight of a satellite that can communicate with a gateway, it can still be covered by means of inter-satellite transmissions.

In sharp contrast to coverage, the throughput performance of these systems differ from each other significantly, mainly due to the number of LEO satellites deployed in the orbits. The peak throughput of 2.66 [Tbps] is forecast to be achieved in the case of Telesat by using 20 [Gbps] optical ISLs and 42 ground stations, while SpaceX system is estimated to achieve 23.7 [Tbps] by using 20 [Gbps] optical ISLs and 123 ground stations.[3] The throughput significantly reduces for SpaceX's system if the ISLs are not used. In par- ticular, more than 50% decrease in the SpaceX's system throughput is observed without inter-satellite communications, which is only around 5% for the Telesat's system using 50 gateways. In addition, Telesat's constellation design appears to be significantly more efficient than SpaceX's one, achieving almost two times more capacity per satellite and utilizing 58.8% of per-satellite capacity as opposed to very low utilization rate of 25.1% in the case of SpaceX.

It is worthwhile to elaborate further on ISLs and the role that they will play in the de- sign of new generation LEO satellite network architectures since they are expected to be an integral part of the space segment of these systems [48, 96]. In the ground-based network architecture not having any ISLs, the logical topology of an LEO satellite network is in the form of many parallel two-hop relay channels, with LEO satellites functioning as the relay connecting gateways and consumer-premise equipment. In this form, a satellite is required to be in the line-of-sight of both the gateway and consumer-premise equipment to provide connectivity. On the other hand, availability of ISLs in the space-based network architec- ture enables multi-hopping among satellites in space in forward and backward directions, and thereby eliminates the former stringent simultaneous line-of-sight requirement. One estimate indicates that deploying ISLs in space leads to almost 60% reduction in the num- ber of gateways without any reduction in the achievable throughput [48]. Although there is not enough data available for the LEO satellite gateway costs, the current gateways for GEO satellite systems range from USD 1M to USD 2M [95]. These indicative cost figures

[3] These are the throughput values in the forward direction by considering gateway-to-satellite and satellite-to-user links.

Table 5: Comparison of SpaceX's and Telesat's big-LEO Constellations

	Orbital Configuration			Payload	Spectrum Usage	Ground Segment
	# of Satellites	Altitude	Inclination			
Telesat	117	1000–1200 km	99.5° and 37.4° (2 sets of orbits)	Digital with phased array antennas	Ka-band	Several gateways and many user equipment
SpaceX 1325	4425	1110– km	53° 81° (5 – sets of orbits)	Digital with phased array antennas	Ka- and Ku-band	Many gateways and many user equipment

and the sheer number of expected gateways that will be deployed on the ground in new LEO network architectures suggest that the deployment costs can decrease significantly if ISL links are used to provide inter-satellite connections. An important challenge in deploying ISLs is the time-varying network topology of LEO satellite systems, which will trigger frequent updates in the routing table at the network layer [96].

An important enabler for achieving the target rates as high as 38 [Gbps] per satellite in new LEO satellite systems is the the use of high frequency bands for transmissions. The new big-LEO constellations aim to use Ka-Ku bands, even with some planning to use V bands as well. Higher frequency bands are, however, more vulnerable to rain fade defined as the absorption of the electromagnetic waves by rain, snow and ice. Hence, advanced techniques in adaptive coding and signal modulation must be developed to alleviate these channel impairments. To this end, a variable coding modulated OFDM (VCM-OFDM) system is developed in [5]. Filtered OFDM waveforms are used to suppress the side lobes of conventional OFDM signals to provide better protection against Doppler shifts. The detailed system-level simulations indicate that the proposed VCM-OFDM approach is robust in the presence of Doppler shifts and provides 43% increase in the system throughput when compared with the current standard DVB-S2 based LEO satellite networks. An alternative approach to the one proposed in [5] is to have a CDMA-based LEO satellite network with continuous wave pilot carriers to simplify code acquisition, effectively compensating extremely high Doppler shifts [102].[4] Although both approaches presented in [5, 102] demonstrate promising performance in combating adverse effects of extreme Doppler shifts in LEO satellite communications systems, the one based on OFDM will be backward compatible with the 5G standards, and hence will be easier to integrate with the existing terrestrial communications systems.

[4] Some capacity enhancement and power control techniques for CDMA-based LEO satellite systems are presented in [103, 104]. These papers are not within the scope of our review since they focus on specific transmission strategies tailored for the earlier generation LEO satellite communications systems.

4.3 Key Findings and Future Research

The following is a list of key findings regarding LEO satellite systems and potential future research directions that are envisaged to play important roles for realizing the full scale of their benefits.

- The development of a principled approach to optimize the deployment of space and ground segments in new generation LEO satellite systems presents great potential to maximize coverage and data throughput as well as minimizing deployment costs. The new generation LEO satellite systems will be massive network infrastructures consisting of thousands of satellites in the space and hundreds of ground stations on the Earth. The existing approaches in the literature depend on some heuristic tech- niques such as genetic algorithm, which are not guaranteed to lead to the optimum deployment or to scale well with the system size.

- The development of a practical framework to optimize the distribution of network algorithms between ground and space segments in new generation LEO satellite sys- tems is critical to fully utilize the benefits that come with digital satellite payloads. Digital payloads will enable implementation of advanced physical layer and net- working algorithms (e.g., phased-array beam steering, modulation, demodulation, routing and congestion control) on-board in new generation LEO satellite systems. While it is more advantageous to run some algorithms on-board to adapt to the rapid changes in the propagation environment, some algorithms may need to run at a ground control center (i.e., network management center) having a general overview of the big-LEO constellation, gateways and user demands. However, there is cur- rently no systematic approach to determine where to place network algorithms in new generation LEO satellite systems.

- The development of an elaborate beam hopping and steering design to handover the users among the satellites seamlessly has the potential to increase system ca- pacity and user experience significantly in new generation LEO satellite systems. High satellite mobility will trigger frequent handovers in these systems. Usual LEO satellite speeds at 5 to 10 km/s will lead to a handover in less than a minute for a typical LEO spot-beam having radius 450 kilometers. Beam footprints with smaller radii will increase frequency re-use and system capacity but at the expense of more frequent handovers. An elaborate beam-hopping and steering design will increase the dwell time of users within a satellite’s coverage area, and thereby alleviate the frequent handover problem. The existing approaches either focus on hybrid spot-beam coverage without hopping among the beams or beam steering and radio resource management without considering handover frequency as an optimization objective.

5 Transceiver Design for combating non-linearity and interference for Satellite Non-Terrestrial Networks (NTN)

3GPP is now developing a new standard to incorporate satellites NTN to augment terres- trial 5G networks. The integrated satellite and terrestrial architectures will significantly change the role of satellite communications in the near future and can lead to manifold advantages, such as coverage extension and additional communication link among many others.

However, NTN integration also faces many unique challenges related to the unique characteristics of Satellite communications, such as high Doppler shift, high non-linearity due to high transmission power, high attenuation and complex interference scenarios. Though there are some existing solutions for Satellite communications to combat these channel impairments, they are designed based on different transmission schemes and can- not work well for 5G waveforms and protocols. How to redesign, optimise and adapt advanced 5G waveforms and key technologies originally designed for terrestrial networks, such as massive MIMO, OFDM for Satellite NTN, becomes critical. Moreover, the wide coverage area and broadcasting nature of the satellites also causes information security is- sues in the satellite communication systems. How to design security and signal processing algorithms for NTN is also very important.

In this section, we will review the existing state of the art algorithms for combating non-linearity, interference and summarize the key findings and identify key challenges and future research directions for redesigning, optimising and adapting 5G waveforms and protocols in NTN.

5.1 Machine Learning for Combating Nonlinear Effects of Satel- lite Amplifiers

Satellite communications play an important role in extending cellular networks to rural and hard-to-reach areas. Due to the nature of long distance transmission, satellite com- munication systems usually operate at a high power range. The amplifier is a primary component in the transmitter side, aiming to adjust an input signal in a low power range to a higher power output signal in a linear way to meet the transmission requirements. However, physical devices can only achieve the ideal linear amplification over a limited range. As shown in Fig. 19, when amplifiers operate at the medium- and high-power sig- nal levels, the outputs may show a degraded performance with nonlinear characteristics.

The nonlinearity is an inherent characteristic of amplifiers and is one of the most challenging problems in satellite communications. Typically, the nonlinearity will result in adjacent channel leakage, degraded error performance, and force the transmitter to reduce its transmission power into a more linear but less power-efficient region [105]. To tackle the nonlinear effects and ensure an efficient power amplification, many methods including classical signal processing approaches as well as state-of-the-art machine learning (ML) techniques have been proposed. In the following, we review three main categories of the existing approaches, and their relative positions are shown in Fig. 18.

- *Post-Compensator*: A post-compensator is widely used to alleviate the nonlinearity by compensating the received signals to counteract the nonlinearity. Tradition-

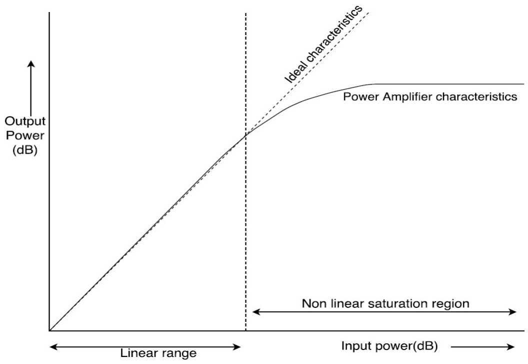


Figure 17: High power amplifier characteristics.

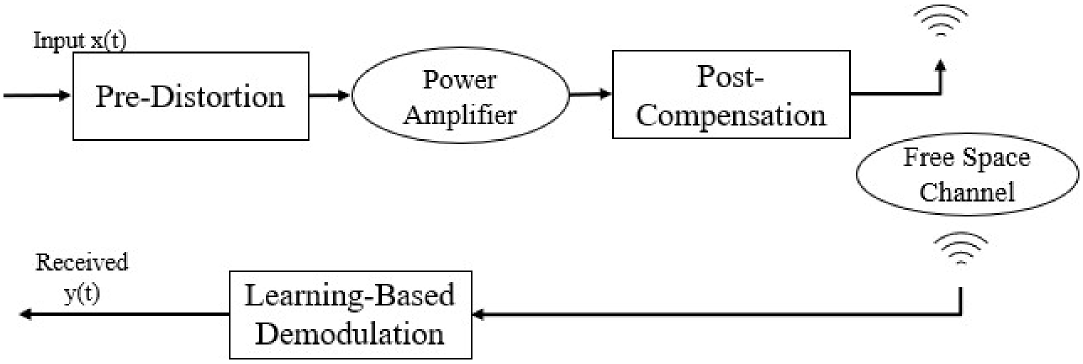


Figure 18: A typical satellite communication system with a power amplifier.

ally, several methods including inverse function and Volterra equalization were in- troduced in [106] [107]. These methods can mitigate the performance degradation caused by the nonlinear effect along with the interference cancellation. Despite their effectiveness, these methods are not suitable in practice due to high computational complexity. To handle this problem, the authors in [108] proposed a low-complexity nonlinearity post compensation technique in satellite communication systems. To reduce the complexity, a look-up table (LTU) with an error detector was developed to store the computed amplitude and angle errors. The LTU requests a training process to obtain the compensation values, and the nonlinearity can be alleviated by applying the compensation amplitude and angle stored in the table. Never- theless, the post compensator has its inherent drawback, i.e., the input signals for post-compensator have already been distorted by the satellite wireless channel and thermal noise. Therefore, an input may not lead to a clean output even with a highly effective compensator. In this case, a more appealing method is to add one pre-distortion component before the power amplifier to pre-process the input signal so that a cleaner and near-linear amplified output can be produced.

- *Pre-distortion*: The basic principles and structures of pre-distortion have been introduced in [109] [110]. In [109], the authors introduced a direct learning architecture by inverting a nonlinear amplifier model. A better indirect learning model was developed in [110]. Once the learning model was established, it can be directly used as the pre-distortion component. However, these methods adopts multiple polynomial terms, resulting in a high implementation cost. In [111], the authors proposed a ML model for the pre-distortion design. By leveraging gradient descent in optimising the cost function, the proposed model can achieve a better error performance given the well-trained model coefficients. However, this method does not function well when the nonlinear effect becomes severe. To overcome this problem, the authors in [112] proposed a novel method based on sparse Bayesian learning (SBL), which can esti- mate the sparse model coefficients from the Bayesian perspective. Comparing with the conventional ML methods, the number of model parameters and training sam- ples can be significantly reduced without sacrificing its error performance. However, the model parameters can be easily impacted by the power amplifier noise. This drawback has been recently addressed in [105], where a neural network (NN) based method was proposed to improve both adjacent channel leak and bit-error rate. In addition, the work in [113] extended the conventional NN to a deep NN (DNN) to construct the pre-distortion component, which outperforms the NN based methond in [105] in terms of the error performance and features a faster convergence rate during the training stage.

- *Learning-Based Demodulation*: Rather than reshaping signals in post-compensator and pre-distortion, the learning-based demodulation aims to leverage the ML tech- niques to alleviate nonlineariry in demodulation process at the receiver. The authors in [114] proposed a support vector machine (SVM) based M-PSK detector to mit- igate nonlinear phase noise. Without any prior information, SVM can learn and capture the properties of the nonlinearity from training data. However, the SVM is only a binary classifier, and multiple SVMs should be allocated for high-order quadrature amplitude modulation (QAM). To address this challenge, the authors in [115] proposed distance-weighted k-nearest neighbors (DW-KNN) based detector to classify the received noisy signal with a high accuracy.

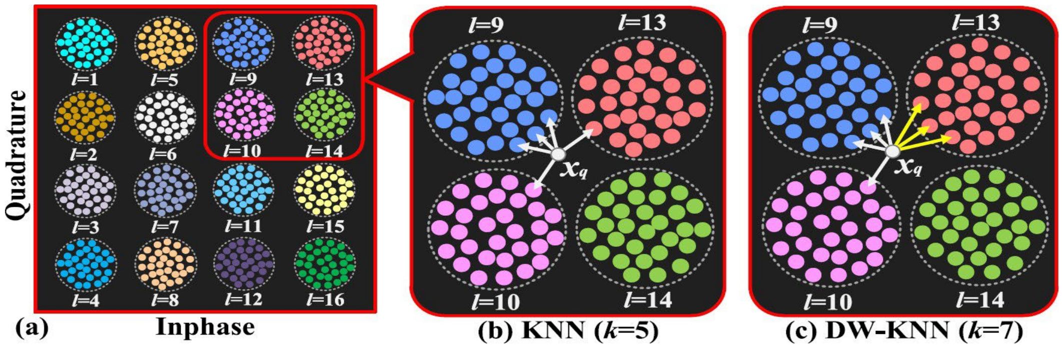


Figure 19: (a) 16QAM constellations with the corresponding labels; (b) an example of one testing data point x_q detected by KNN (k =5); (c) the testing data point x_q detected by DW-KNN (k = 7) [115].

5.2 Machine Learning Based Receiver of 5G NTN MIMO Sys- tem

5G NTN communications systems will use massive MIMO Line-of-Sight (m-MIMO LOS) for maximising the number of connections and reliability [116]. 5G NTN will integrate 5G New Radio (NR) with Very High Throughput Satellite Systems (VHTS). As the number of antennas, the order of the constellation and the number of users increase, the design of receivers which can efficiently remove interference and have low complexity is very challenging. There are three types of interference 1) inter-carrier interference due to the Doppler shift caused by the movement of VHTS satellite; 2) inter-beam interference due to aggressive frequency reuse; and 3) multi-user interference in Line of Sight (LoS) Multiuser (MU) MIMO systems for satellite communications. MIMO receivers can be classified into two categories: classical and machine-learning receivers. As the complexity of the optimal classical detector such maximum likelihood (ML) receivers and sphere decoding based receivers [117], increase exponentially with the constellation size and the number of users and the number of antennas, they are not suitable in practical applications.

To reduce the computational complexity compared to the optimal ML receivers, it- erative classical receivers consisting of non-iterative (e.g. minimum-mean-square-error (MMSE)) and iterative receivers have been proposed [118] to cancel inter-beam or multi- user interference. Among the low-complexity classical detectors, the iterative MMSE with successive interference cancellation (SIC), MMSE-SIC detectors, achieve the best BER performance [119]. However, the computational complexity of the MMSE-SIC re- ceivers increases polynomially with the number of antennas due to performing the MMSE matrix inversion operation in every iteration. The polynomial computational complex- ity quickly becomes prohibitive with increasing number of antennas. To address this issue, the matched filter, parallel interference cancellation (PIC) scheme and the decision statistics combining (DSC) scheme are integrated in the PIC-DSC detectors [120]. They first perform the matched filtering, i.e. multiplying the received signals by the conjugate transpose of the channel matrix, and then estimate the interfering symbols using the DSC scheme. Finally, the PIC scheme is used to subtract these symbols from the received sig- nals in a parallel manner to recover the desired symbols. As shown in [120], the PIC-DSC receivers have a similar performance to the MMSE-SIC receivers and a linear computa- tional complexity in terms of the numbers of users and antennas. The same iterative PIC-DSC receiver technique has been to provide symbol feedback, seen by channel esti- mator function as additional virtual pilots. This results in much better channel estimate that in turn further improves symbol detection [121,122]. These additional pilots are used to track inter-carrier interference that are caused by a high Doppler spread.

More recently, machine-learning receiver based on Bayesian rule, referred to as Bayesian M-MIMO receivers have been proposed to significantly improve the performance compared to the classical detectors by incorporating detection probability measures when estimating symbols from the received signals [123]. The Bayesian framework is used to satisfy the maximum a posteriori (MAP) criterion for MIMO detection by using the Bayesian rule. The optimal Bayesian receiver in terms of the BER performance use the Bayesian frame- work with the MMSE scheme as their inter-beam or multi-user interference suppressor, referred to as expectation propagation (EP) detectors [124–126] and including our work in [127]. Despite a significant BER performance improvement compared to the MMSE detectors, the EP detectors perform MMSE matrix inversion operation in every iteration resulting in a polynomial growth of the computational complexity with the number of receiver antennas. Another class of low-complexity Bayesian based receivers, which in- cludes the approximate message passing (AMP) [128] and the approximate EP [126, 127] detectors, has been proposed to avoid the matrix inversion operations. However, the BER achieved by these detectors is much worse than an accurate EP receiver, which is a near optimal receiver. In our recent research we have shown that combining Bayesian framework with PIC-DSC for M-MIMO receivers results in EP performance close to an EP receiver [129]. Unfortunately, all the above advance machine-learning based receivers deal only with M-MIMO non-LOS receivers. To the best of authors' knowledge, there is no Bayesian based M-MIMO LOS receiver able to achieve a near optimal BER with a linear computational complexity that simultaneously address Doppler spread, inter-beam or multi-user interference.

We also note that recently deep learning (DL) in the popular form of deep neural net- works (DNNs) has been used to perform interference mitigation and channel estimations. This leads to a problem in the lack of transparency and trust in logic decisions as com- pared to traditional mathematical model-based optimization. NNs with multiple layers cannot explain the essential features that influence actions or the impact of data bias on the uncertainty of outputs [130]. Furthermore when very good mathematical models with clear inputs and outputs relationships such as the one used in channel estimations, multi- user and inter-carrier interference mitigation due to multiple beams and Doppler spread are well known, Bayesian inference outperforms Deep learning [130], further justifying the feasibility of the research direction in machine-learning for developing m-MIMO LOS receivers for 5G NTN.

5.3 Physical Layer Security of 5G NTN MIMO System

Physical layer security is an important consideration in beyond-5G satellite NTN. This is because the large number of transmission beams between the ground stations, satellites, and users are vulnerable to security and privacy attacks due to the shared wireless medium and aggressive frequency reuse needed to achieve high throughput gains.

Secret key generation based on physical layer characteristics of wireless channels has become increasingly popular due to its information-theoretical security and lightweight properties compared with cryptographic encryption [131]. In this security strategy, sym- metric keys can be generated between legitimate users through channel reciprocity in a short duration, whereas potential eavesdroppers are assumed to have no useful knowl- edge about the key because their channels are spatially decorrelated with those of the legitimate users.

Most research works in this area have focused on improving the performance of key generation by refining existing algorithms or considering different wireless channel sce- narios. Specifically, the effect of multiple antennas on the key generation rate (KGR) was investigated in [132]. In [133], the amplitude-phase joint quantization for multi-carrier systems was proposed to make better use of channel state information (CSI). Schemes for key sharing in groups [134] and untrusted relay networks [135] have also been presented. However, there is a lack of attention to the security of key generation method itself. From the perspective of eavesdroppers, the security of physical layer secret keys is generally overestimated. This is because malicious jamming may destroy the key symmetry during key establishment. Given the pilot sequences, malicious nodes can impersonate legiti- mate users using spoofing attacks. In our previous work, we have investigated security vulnerabilities in LoS channels where the propagation conditions are known to the eaves- dropper [136]. We will build on our existing work to develop new low-latency wireless key generation protocols for satellite NTN to minimize channel correlations between users in LoS MU-MIMO channels. We will consider the impact of physical-layer security and privacy attacks such as eavesdropping and covertness [137] in satellite NTN. Our work aims to establish fundamental performance trade-offs between the latency, reliability and security requirements in beyond-5G satellite NTN.

Another promising approach to improve physical-layer security is to harness interfer- ence from friendly external jammers to assist the legitimate users [138, 139]. In [140], friendly jamming strategies were proposed to improve an area-based metric known as the jamming coverage to minimize the secrecy outage probability for unknown eavesdropper location. In [141], the outage probability (OP) of the legitimate receiver and the inter- cept probability (IP) of multiple eavesdroppers were derived to analyze the impact of the jamming power on the security-reliability tradeoff of friendly jamming. In [142], a novel algorithm to match every source and destination pair with a jammer based on match- ing theory. In [143], the secrecy performance of three artificial-noise-aided transmission schemes are examined.

Recently, the use of unmanned aerial vehicles (UAVs) has been proposed to improve the coverage and security of wireless networks by exploiting the characteristics of their flexible deployment and strong line-of-sight (LoS) links compared to conventional terres- trial networks [144]. In [145], the optimal UAV deployment was investigated to maximize the number of ground users served by a UAV subject to quality-of-service (QoS) con- straints. The physical-layer security of UAV-assisted networks has also attracted growing interest. In [146], the authors jointly optimized the UAV transmit power and trajectory to minimize the OP of a UAV relay. Our initial works have established efficient opti- mization frameworks to minimize the intercept probability for line-of-sight (LoS) UAV communication systems when the locations of eavesdroppers are unknown [147]. In [148], we analyzed the impact of low-latency edge computing networks to improve the security of UAV communications.

5.4 Key Findings and Future Research

The following is a list of key findings regarding combating nonlinear effects and physical layer security:

- For satellite communications, nonlinearity is a crucial factor affecting system performance in the physical layer. Digital pre-distortion is the most popular approach to mitigate nonlinearity. Conventional digital pre-distortion has been doing a good job when nonlinear effects are moderate. The existing methods of combating nonlinearity are mostly focused on single-carrier transmission.

- Learning-based methods can improve the performance of digital pre-distortion, especially in the severe nonlinear cases. Machine learning based demodulation scheme cannot alleviate nonlinearity directly in the physical layer, but it is able to effectively reduce the influence of nonlinearity on the received signals by learning nonlinearity's impact pattern.

- Most works about learning-based pre-distortion are focused on single-carrier systems. OFDM systems suffer from a more complicated nonlinearity problem. OFDM systems have a higher peak to average power (PAPR), which shifts the input signal to the nonlinear region of the power amplifier. Existing pre-distortion and post-compensator algorithms cannot effectively combat the severe non-linearity effects in OFDM systems. Machine learning offers a promising and powerful tools to address the challenging nonlinearity problem in OFDM systems.

- Satellite systems are equipped with a large number of antennas communicating with multiple users each equipped with multiple antennas. How to combat the nonlinearity, interference and design the optimal transceiver for such systems becomes very complicated due to the complex distortion as well as inter-carrier, inter-user and inter-beam interference caused by non-linearity, MIMO and Doppler shifts. The existing literature did not consider all three types of interference.

- The problem of developing a low complexity receiver to minimise inter-carrier, inter-beam and multiuser interference has yet to be addressed. How to develop a machine learning based algorithm for such systems is an open and important research topic.
- For secret key generation, previous works have focused on terrestrial networks with ground users and no works have considered wireless key generation in NTN. For friendly jamming, initial works have considered LoS UAV links to ground users but few works have considered satellite friendly jamming in NTN.
- The physical-layer security performance of satellite NTN with inter-user interference between the satellites and users, and inter-beam interference between the ground stations and satellites has not been addressed in the literature.

6 Edge Computing for Satellite Systems

As outlined in the 3GPP Release 16, 5G satellite access is important for the evolution of the 5G network. The future terrestrial-satellite network integrated with 5G are en- visaged to have low delay, high bandwidth, and ubiquitous coverage [149]. However, the traditional satellite communication systems only have limited on-board computational power [150]. To provide for more complex future services, such as dynamic channel esti- mation and forecasting used for routing and beamforming design, today's satellites need to transmit a massive amount data to ground computing centers and wait for the solu- tions calculated and transmitted by ground co mputing centers, which takes up a large amount of bandwidth and causes a high delay [151]. In order to solve the delay and bandwidth problems, the integrated 5G network uses edge computing techniques to place part of the computing resources at the edge of the network. The advantages of a 5G satellite network edge learning system over the existing satellite network architecture are illustrated by two examples. The 5G satellite network scenario is shown in Figure 20. The meaning of each time cost is explained in Table 6. The user is located in a remote area and wants to connect to cloud and process some data. Using the traditional satellite communication systems, the user needs to transmit the data to the ground data center to process the data via the satellite network. So the time delay of users' access to data center is $T1 + T2 + T3 + T4$. However, if the satellite edge computing network is used, the total time delay is only T 1. There is a ongoing study on edge computing in 5G Core network (5GC) in 3GPP Release 17 Technical Report 23.748.

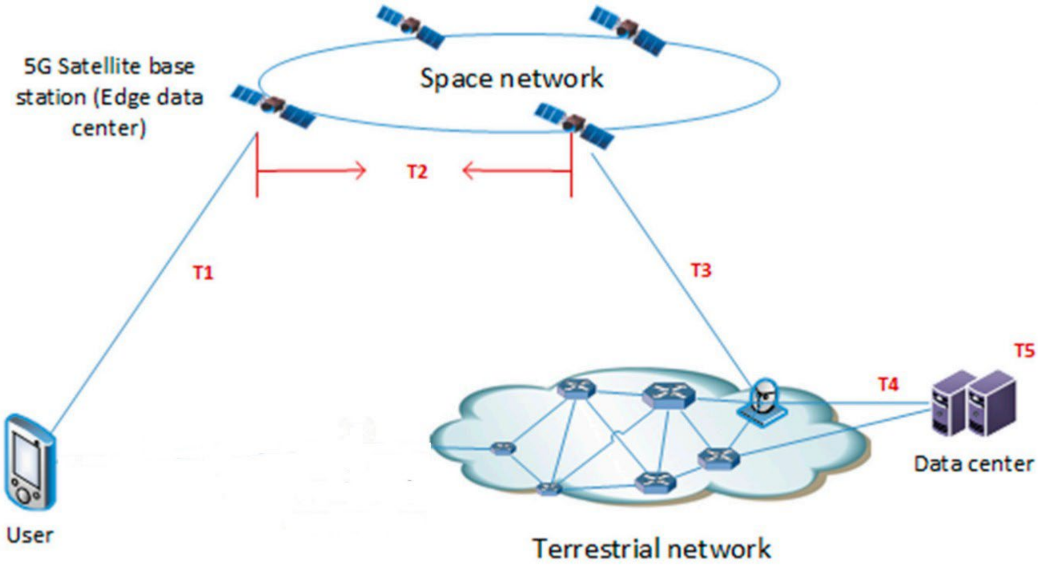


Figure 20: A 5G satellite network scenario.

6.1 Brief Review on Edge Computing

In this subsection, we provide an overview of edge computing, including the strong mo- tivations of moving computing to the edge, the architecture of edge computing, and one example of edge computing platforms, e.g., the cloudlet.

Table 6: The meaning of each time cost.

T1	The time cost of uploading data
T2	The time cost of forwarding data among intersatellite link
T3	The time cost of downloading data
T4	The time cost of forwarding data to data center
T5	The time cost of processing data

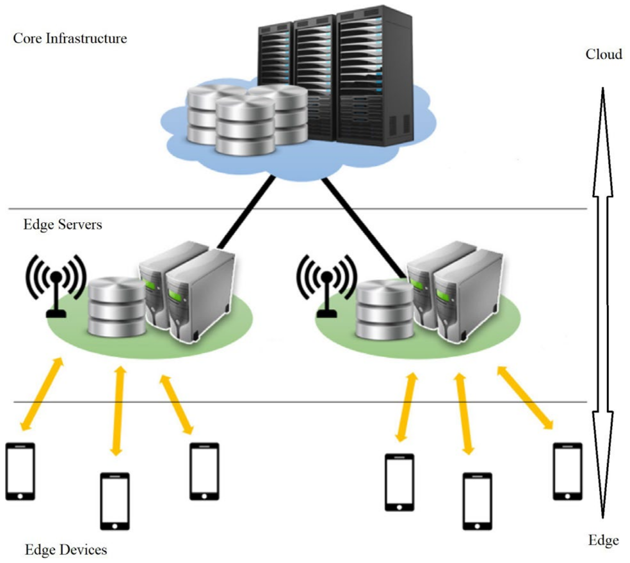


Figure 21: The general edge computing architecture

The advantages of edge computing over cloud computing are numerous. First, it can strike a good trade-off between the AI-model complexity and the model-training speed. Additionally, thanks to its close distance to data sources, edge computing is able to process real-time data, avoiding the over the top propagation delay and the network traffic congestion caused by uploading data to the cloud. Moreover, the proximity of edge computing offers an extra advantage of location-and-context awareness. As a result, edge computing can bolster a wide range of AI models to enable a large series of mission-critical applications, such as autonomous driving, rescue-operation robots, disaster avoidance, and fast industrial control.

A general architecture of edge computing is demonstrated in Fig. 21. It has a three- layer functional structure: core infrastructure, edge servers, and edge devices. The first layer is core infrastructure layer, which provides the core network access and enables the global model consensus and management functions for mobile edge devices. The second layer is edge server layer, which belongs to the infrastructure provider. Usually, the multi-tenant virtualization infrastructure is installed in these edge servers providing the virtualized and multiple management services. Local models are trained at this layer. Besides, the edge computing infrastructure enables the connections among edge devices, edge servers, and the core infrastructure via wireless network, data center network and the Internet. The third layer is edge device layer, where data is generated to feed into the edge computing platform.

Learning with cloudlet is a typical example of edge computing. Cloudlet can be regarded as the edge of the Internet and is proposed to resolve the issue of excessive delay in end-to-end communication between a mobile device and its associated cloud [152]. The

Table 7: Existing work on resource allocation in edge computing

Work	Contributions on resource allocation
Liu et al. [153]	Tradeoff on energy, latency, and offloading costs
Wang et al. [154]	ENORM: Resource management framework of edge node:
Tan et al. [155]	Resources allocation and Caching
You et al. [156]	Resource allocation for MECO systems
Xu et al. [157]	Enhanced resource management algorithm for online learning
Liu et al. [158]	Bandwidth-based partitioning scheme

cloudlet is a small-scale cloud Data Center (DC) and is designed to offer strong processing capacities to smart devices with low communication latency.

6.2 Resources Management And Allocation in Edge Computing

In decentralized edge computing environment, resource must be allocated, such as pro- cessor, disk, and network bandwidth for distributed data processing. Since edge devices may have limited resources including computing, storage, and networking I/O, resource allocation must be performed based on both existing available resources and performance constraints. Specifically, resource allocation is performed under multiple conditions, in- cluding resource usage quota, power and energy consumption budget, and latency.

We list some research work on resource allocation in edge computing environment in Table 7.

Liu et al. [153] tried to tradeoff between energy consumption, execution delay and of- floading cost, and proposed an optimization strategy for optimizing these three objectives simultaneously. Their simulation experiments show that the joint optimization strategy can guarantee better quality of service.

Since the data arrival pattern and deadline for data processing vary significantly in different edge computing scenarios, it is not feasible to formulate general resource al- location mechanism in edge-cloud collaboration environment. You et al. [158] studied the energy-saving resource management strategy of asynchronous mobile-edge computa- tion offloading (MECO) systems. The best data partitioning and time division policy is derived by analyzing the general arrival data series, and then the total mobile energy consumption is minimized by using the block coordinate descent approach.

Similarly, Wang et al. [154] proposed and developed the edge node resource manage- ment framework, namely, ENORM. They proposed a new configuration and deployment mechanism for linking communication between edge nodes and the cloud data center such that ENORM can provide offloaded workloads for edge nodes. Moreover, ENORM inte- grates low overhead and dynamic autoextension mechanism to add or remove resources to manage workloads on edge nodes effectively. They validated the feasibility of ENORM through context-sensitive and delaysensitive online gaming use cases and the results show that ENORM can reduce application service latency up to 20% to 80% and reduce the frequency of data transmission and communication between edge nodes and the cloud up to 95%.

Currently, the networked systems are increasingly prone to be heterogeneous in terms of hardware configuration, software stack, networking media, and application domains. Specifically, data volume, data producing speed and service quality are highly diverse in edge-cloud collaboration environment. Such heterogeneity poses lots of challenges, such as how to address the shortage of mobile device resources, and how to tradeoff between the limited computing power and energy constraints of mobile nodes. Tan at al. [155] designed a virtual and fully duplex small scale cellular network framework based on caching heterogeneous services in edge computing. They proposed a novel resource allocation scheme that not only considers the caching mechanism, but also adopts fully duplex communication. Moreover, the proposed scheme also considers user correlation, power control, caching, computation offloading strategy, and resource allocation at the same time.

In mobile computing environment, energy consumption is the key concern for resource allocation and computing performance maximization. Researchers proposed energy sav- ing approaches for resource allocation of single user and multiuser mobile edge computing offloading systems (MECO). However, these existing works focus on the design of complex algorithms rather than the design of optimal resource allocation strategy. You et al. [156] investigated the resource allocation of multi user MECO systems based on time division multiple access (TDMA) and orthogonal frequency division multiple access (OFDMA) and consider cases with infinite or limited cloud computing capabilities. For TDMA mo- bile edge computing offloading systems with infinite cloud computing capabilities, they propose the resource allocation strategy by redefining the offloading priority function and modifying the previous threshold policy and then propose a low complexity sub-optimal resource allocation algorithm based on the approximate offloading priority. In other hand, for OFDMA mobile edge computing offloading systems with unlimited cloud computing capabilities, they solve the resource allocation problem as a mixed integer optimization problem and the prioritized TDMA strategy is used to optimize resource allocation, which includes: (1) translating the OFDMA resource allocation problem into a corresponding part of the TDMA, (2) determining the initial resource allocation and offloading data by defining an average offloading priority function, (3) assigning the sub channels according to the offloading order, and (4) adjusting the allocation of the offloading data on the sub channels. Simulation experiments show that this resource allocation strategy can ap- proach optimal performance. However, the proposed approach also has some shortcomings in that they assume that: (1) the processed data can be processed separately, (2) each mobile device can perform local computation and incoming workload offloading at the same time, and (3) the edge cloud has a complete understanding of energy consumption in the local computing devices, channel gain and fairness factors of all users.

Although currently renewable energy is used to power the mobile edge computing ca- pabilities, the intermittent and unpredictable nature of renewable energy poses a huge challenge for high quality computation offloading services. To solve this problem, Xu et al [157] defined this problem as a Markov decision process and proposed an efficient online resource-based reinforcement resource management algorithm, which can reduce system service latency and operating costs by real-time learning of the best strategies for dynamic job offloading and edge server provisioning. Unlike traditional reinforcement learning algorithms, the proposed online learning algorithm achieves higher learning rate and runtime performance through decomposition value iteration and reinforcement learn- ing. The simulation results show that the system cost of the online learning algorithm is much lower than that of the compared schemes. In addition, the results also show that the proposed approach can save more power especially when the network connection is deteriorating.

In summary, all these literature only considered edge servers with fixed locations, the dynamically changing nature of LEO satellites and the satellite constellation poses a huge challenge for high quality resource allocation and computation offloading services. The power generation prediction of renewable energy source for the edge data center is not considered. The channel estimation errors and feedback delay can limit the performance gain. However, the channel condition prediction of the wireless link for wireless network is not considered in these papers. There is a need to develop a resource management algo- rithm for edge computing for 5G NTN, which consider power generation by the on-board solar cell array, power consumption of on-board data processing tasks and communication tasks, and on-board processing capability.

6.3 Edge Computing for 5G NTN

In [151], the authors proposed a 5G satellite edge computing framework (5GsatEC), which aims to reduce delay and expand the network coverage. This framework consists of an em- bedded hardware platform and edge computing microservices in the satellites. To increase the flexibility of the framework in complex scenarios, the authors unified the resource management of the central processing unit (CPU), graphics processing unit (GPU), and field-programmable gate array (FPGA); they divided the services into three types: system services, basic services, and user services. However, only simple simulations were carried out and no practical experiments have been conducted. In [159], the authors proposed a satellite IoT edge intelligent computing architecture based on the latest development of edge computing and deep learning. Among them, the satellite IoT edge computing and the distributed satellite IoT intelligent computing architecture were described in detail, which can be used as the edge learning methods to train the machine learning model in a distributed manner. By testing the performance of different neural network models, it was observed that a lightweight neural network model is suitable for satellite IoT scenarios.

Convolutional neural networks (CNNs) have been successfully applied to many com- puter vision applications. However, the current state-of-the-art CNN models require massive computational power and memory capacity. This has limited their application in the edge AI platforms. As a result, researchers have been seeking ways to conduct computation approximation and model compression in deep neural networks, while pre- serving the performance of machine learning. Generally speaking, this approach can be divided into four categories: parameter pruning and sharing, low-rank factorisation, trans- ferred/compact convolutional filters and knowledge distillation. In the future, we plan to use low-rank factorisation methods to develop lightweight neural network model for satel- lite edge learning. Low-rank factorisation methods use matrix decomposition techniques to search through the model parameters and find the most influential model parameters in the deep CNNs. Note that the convolution kernels can be viewed as a 2D matrix for the fully-connected layer. Hence, performing a low-rank factorisation on such matrix is useful to identify the major components, and remove the redundancy in the model parameters.

6.4 Key Findings and Future Research

The following is a list of key findings regarding edge computing for 5G NTN.

- Although edge computing has been considered for 5G terrestrial networks, it has not been applied in 5G NTN. These existing papers only considered resource man- agement problems for edge servers with fixed locations, the dynamically changing nature of LEO satellites and the satellite constellation poses a huge challenge for high quality resource allocation and computation offloading services.

- The impact of the shortage and instability of power supplied by arrays of solar cells on edge computing has not been considered in the literature. Power generation prediction of renewable energy sources for the edge data center has also not been considered.

- The channel estimation errors and feedback delay can limit the performance gain. However, the channel condition prediction of the wireless link for wireless network is not considered in these papers.

- There is a need to develop a resource management algorithm for edge computing for 5G NTN, which consider power generation by the on-board solar cell array, power consumption of on-board data processing tasks and communication tasks, and on-board processing capability.

- There is an opportunity to develop offloading schemes to achieve parallel computation between satellites.

- The relative velocity between the transmitter and the receiver introduces a Doppler shift. For a typical LEO satellite at the altitude of 650 kilometers, the Doppler shift varies from -4 kHz to 4 kHz. As LEO satellites have fixed trajectories, so their Doppler shifts are predictable. Doppler shift estimation and prediction at the network edge to pre-compensate the effect of the Doppler shift is an important challenge. Machine learning methods offer a promising approach that has not yet been attempted.

7 FPGA-Based Fault Tolerant SDR Satellite Transponders

Software-defined radio (SDR) is a combination of software and hardware. It allows ra- dio systems to perform Intermediate Frequency (IF) processing by using software while the digital signal processing algorithms, such as filtering, modulation, channel coding, and MIMO processing, can be operated on programmable logic devices (PLD) like field- programmable gate array (FPGA) or digital signal processor (DSP). As many key analog components, which are commonly expensive, have been partially replaced by digital algo- rithms, SDR can perform the same functionalities as analog devices at a lower cost, while providing better stability.

At present, FPGAs are widely used for data flow in SDR applications to achieve configurability and flexibility. In [160], the universal software radio peripheral (USRP) was used to form a modular FPGA-based SDR for CubeSats. The USRP has a Xilinx Spartan-3A 1400 FPGA to up-converts output data to an IF band and down-converts in- put data through a series of cascaded integrator-comb (CIC) and half-band filters. Some other researchers tried to fully use the featured pipeline and parallel design abilities of FPGAs to improve the system efficiency and adjust the power consumption. The pro- grammable ultra-lightweight system adaptable radio (PULSAR) developed by [161] used an Actel PriAsuc3 FPGA in small satellite communication systems. Compared to the use of multiple processors or multi-core processors, FPGA has the ability to perform parallel and pipelined functions to achieve a much more efficient use of the clock so that a much lower overall clock speed is required and the system size, cost, and power consumption can be limited as well. The authors focused on minimising the analog and RF components by converting the functions into the digital format. All signal processing algorithms were designed using hardware description language (HDL) and handled inside the FPGA.

SDR is becoming more and more popular in the space application as the problem of components replacement and update can be easily solved by using SDR systems. The controllable power consumption and high flexibility also make SDRs a better choice when comparing to the traditional radio systems. Currently SDRs are used for both inter- satellite communications and earth-satellite links at a wide range of frequency bands. In [162], the Institute of Space Systems of the German Aerospace Center (DLR) devel- oped a Generic SDR system which can operate on multiple bands with its reconfigurable and small platform. The GSDR has a motherboard based on a Xilinx-7000 System-on- Chip (SoC) baseband processing system and a RF daughterboard containing specific RF circuits, filters, mixer and amplifiers. In [163], NASA's Glenn Research Center made a flight test-bed called SCaN with reconfigurable and reprogrammable SDRs which can op- erate at Ka-band with the RF/antenna systems required for communications. The system applied Harris corporation with space-based digital processing platforms and terrestrial- based SDR technology which is NASA's first space-qualified Ka-band transceiver. The Harris SDR can provide a Ka-band duplex radio link with the tracking data and relay satellite system (TDRSS). The final product is a reprogrammable and highly-capable Ka-band transceiver which can handle data rates greater than 100 Mbps.

As the space environment is highly different from the conditions on earth, it is required to analyze the harsh conditions, which may occur in the space, and test the system in a simulated environment to ensure that it is also able to operate successfully in the space, where the adjustment and repair are hard to conduct. In [164], the researcher listed several space challenges for SDRs including radiation effects, frequency uncertainties, signal fading, reconfigurability time, and interference with adjacent channels. A single event upset (SEU) occurs when a charged sub-atomic particle causes a state change in an electronic component. SEU mitigation techniques were considered, including: triple modular redundancy (TMR) and error detection and correction (EDAC) codes. A de- tailed investigation on system hardening strategies for radiation effects was given by the GSDR system in [162]. Devices with error correction codes, radiation-tolerant compo- nents, aggregate regulation and a distribution system, an overvoltage protection circuit, current-sense monitoring network, and other different fault-tolerant mechanisms were ap- plied to protect specific modules or solve specific problems. In [163], the Harris SDR was tested based on digital and RF performance, random vibration, thermal vacuum, electromagnetic interference, and electromagnetic compatibility. The results can verify that if the SDR met all requirements and if it was capable for operating in the space environment.

Industries and technology companies are developing new products and devices to help researchers conduct 5G and artificial intelligence related SDR developments. Both Xilinx and Intel have hardware optimised FPGA solutions and software defined environments to support the implementation of 5G New Radio networks. In [165], National Instruments (NI) has developed several products focusing on solving particular 5G technical barriers. A massive MIMO Software Architecture with LabVIEW system design software and NI USRP SDRs can be used to build testbeds. NI also provided an SDR platform with a mmWave transceiver system, which allows researchers to modify and optimise real-time over the air mmWave communications.

Cyber security is another emerging aspect in satellite communication systems. Traditional wireless authentication relies on software identity such as password and MAC address which can often be stolen by unauthorised users. It is necessary to adopt addi- tional authentication mechanisms to enhance and secure assets from these attacks. Ra- dio frequency (RF) fingerprinting is one of the promising characteristics which can be uniquely identified with respect to the individual physical transmitters. These RF fin- gerprints are naturally imparted by hardware imperfection during manufacturing within a range of tolerance so that it can be considered as a radiometric signature for an al- ternative authentication. Therefore, by taking advantage of these characteristics, we can distinguish and recognise a specific transmitter by constructing and training a deep neural network. Recently, DARPA has announced the Radio Frequency Machine Learning Sys- tems (RFMLS) program to develop the foundations to apply data-driven ML to the RF spectrum domain [166]. Azarmehr et al introduced a novel approach using phase locking mechanism for device identification based on the imperfections of the RF oscillator of a transmitter [167]. A comprehensive overview of challenges in the RF fingerprinting can be found in [168].

7.1 Key Findings and Future Research

The following is a list of key findings regarding FPGA-based fault tolerant SDR satellite transponders.

- Fault tolerance in the space environment has been studied many times and multiple radiation-resistant devices, correction algorithms, and solutions are provided in the area.

8 Improving End-to-End (E2E) Performance with Software-defined Satellite Network Architectures

SDN has brought flexibility and agility to operators of the terrestrial networks [169]. Espe- cially, it improves the efficiency of network management and optimisation by centralising the control plane. In this section, we discuss how to use software-defined satellite network architectures to improve the E2E performance of the satellite networks.

8.1 Standards for Software-defined Satellite Networks

8.1.1 Digital Video Broadcasting Standards

Legacy Digital Video Broadcasting (DVB) standards [170, 171] allow multiple virtual network operators (VNOs) to share one satellite network by assigning different Internet Protocol (IP) addresses to VNOs. Then, the network sends the packets of different VNOs according to the assigned IP addresses. Based on this fact, the authors of [172] pro- posed a software-defined satellite network architecture that abstracts the DVB-standard- based satellite network as an Openflow switch, which enables operators to program the packet switching and routing via the control plane. However, because VNOs share DVB- standard-based satellite networks at an IP-packet level, the underlying radio resources are out-of-control of each VNO. Without controlling the allocation of radio resources, it is hard to achieve a fine-grained QoS in terms of throughput, delay, and reliability. Therefore, future satellite networks are required to have high flexibility and programma- bility over radio resources that will be shared among operators with heterogeneous service requirements.

Recently, satellite operators started to deploy their proprietary networks to provide heterogeneous services, e.g., mobile broadband services by SpaceX and the Internet of Things (IoT) services by Fleet Space. However, the deployment and operation of proprietary satellite networks are costly. To reduce the cost, network operators can share satellite networks by using SDN architectures and each operator becomes a VNO.

8.1.2 5G Non-Terrestrial Network (NTN)

The success of cellular networks in terrestrial has attracted the attention of the satellite industry. Specifically, the 5th generation (5G) cellular network is standardised with flex- ible configurations of the physical layer [173], e.g., bandwidth and sub-carrier spacing, which could be suitable for supporting services with heterogeneous requirements. The 3rd Generation Partnership Project (3GPP) has started the investigation of using 5G as the satellite communication protocol to provide Internet connections. Such a 5G-based satellite network is referred to as 5G NTN. The authors of [174] implemented a software- defined core network that can be used in a 5G NTN. In addition, 3GPP has summarised potential issues when implementing the radio access network (RAN) of 5G NTNs [175] as well as the possible solutions to these issues [176]. However, there is no available imple- mentation (either in simulation or real-world) of RAN of 5G NTNs in the open literature. As a result, how to establish a software-defined architecture for 5G NTN remains an open challenge.

8.2 Improving End-to-End Performance in 5G NTN

In this project, we will establish a software-defined satellite network architecture for 5G NTN. Based on this architecture, it is possible to realize a full E2E networking concept that has the potential to satisfy the QoS requirements of new application scenarios in 5G, i.e., enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC) [177]. Since high- throughput satellite communication systems have been reviewed in previous sections, in this subsection, we focus on the latency, jitter, reliability, and age-of-information (AoI) in satellite networks.

8.2.1 E2E Performance Analysis

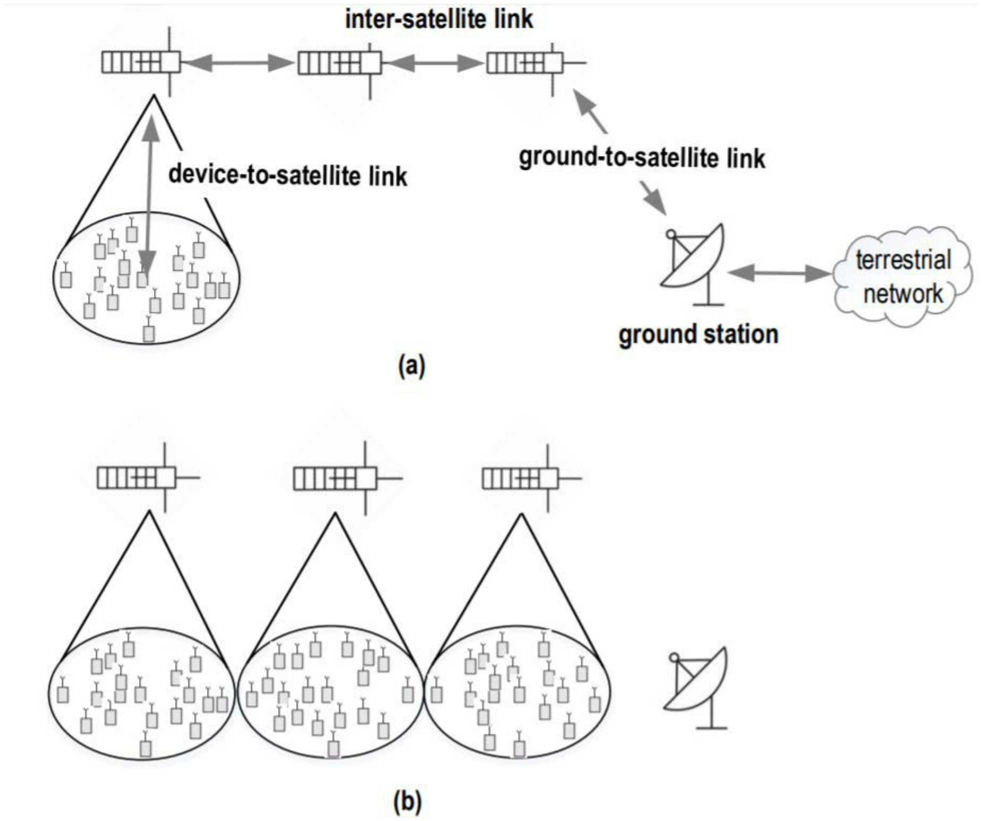


Figure 22: Example of multi-hop relaying satellite network [178]. (a) Only the first satellite receives data from ground. (b) All satellites receive data from ground.

To analyse the performance of a path, one can use the model-based method in queueing theory. As illustrated in Fig. 22, the analysis highly relies on the assumptions on the traffic models, channel models, and the topology of a network. For example, by adopting the Jackson's theorem to the SDN architecture of satellite networks, the average sojourn time of a file that consists of multiple packets was derived in a closed-form expression in [179]. The authors of [178] analysed the average waiting time and the average AoI in a multi-hop satellite network. Both of these works assumed that the queueing networks can be modelled as M/M/1 systems, which may not hold in practice. When theoretical models are not available, a deep learning approach with graph-based neural networks (GNNs) was proposed to predict the mean delay, jitter, and reliability in terrestrial networks with fixed topologies [180]. Such an approach is referred to as model-free method.

8.2.2 Routing Scheme Optimisation

Optimising routing schemes in LEO satellite networks is more challenging than that in terrestrial networks. Since LEO satellites move fast around the earth, the topologies of satellite networks are dynamic. In addition, with thousands of satellites, the scale of the problem is large. Thus, the control plane of a network needs to optimise routing schemes in dynamic large-scale networks.

Intuitively, the E2E delay from one node to another increases with the number of hops between them. One can apply Dijkstra's shortest path first (SPF) algorithm to minimize the number of hops between the two nodes. However, in satellite networks, there are two types of ISLs: intra-plane ISL that connects with the satellites in the same plane and inter-plane ISL that connects satellites from different orbital planes. Since the communications over intra-plane ISL are more stable and easier than inter-plane ISL, the SPF algorithm is not optimal. The authors of [181] developed an algorithm that minimizes the number of inter-plane hops in satellite networks at the cost of more intra- plane hops. By combining GNNs with deep reinforcement learning, the routing scheme can be optimised to achieve better tradeoffs among delay, jitter, and reliability [182]. Furthermore, GNNs have two important features: scalability and transference, which allow the control plane to optimise routing schemes in dynamic large-scale networks [183]. Nevertheless, how to apply GNN-based deep reinforcement learning in satellite networks remains unclear and deserves further investigation.

8.2.3 Learning-based Programmable Scheduler Design

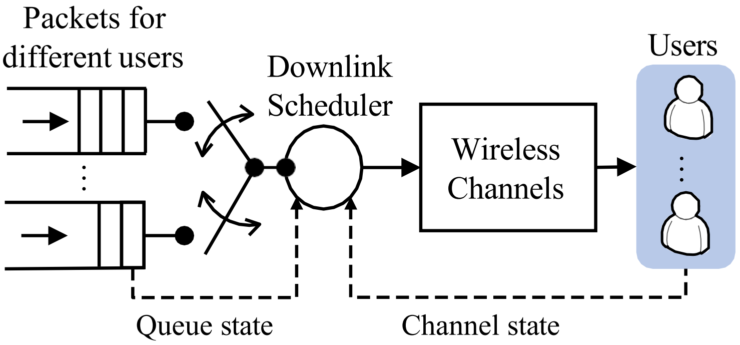


Figure 23: Illustration of a scheduler [184].

Although there are a large number of scheduling policies in the existing literature, none of them can meet the diverse QoS requirements in 5G NTN. Thus, wireless schedulers in satellite networks should be re-designed. As illustrated in Fig. 23, a wireless scheduler is a multi-dimensional function that takes the queue state information (QSI) and the channel state information (CSI) as its input and outputs the amount of resources allocated to different users. We can approximate the scheduling policy by using a DNN and and optimise the parameters of the DNN by using deep reinforcement learning [184]. Such an approach has been implemented in terrestrial scheduler design, where the design objective is to achieve low latency and low jitter with high reliability [184]. The results showed that learning-based schedulers can achieve high reliability than the heuristic schedulers such as the earliest-deadline-first scheduler and the maximum throughput scheduler.

Nevertheless, how to apply the learning-based approach in satellite networks remains an open problem. The major challenges in satellite scheduler design lies in the following aspects: 1) Due to long propagation delay, the estimated CSI is outdated. 2) High Doppler shift in LEO satellite communications leads to significant performance losses in terms of throughput, reliability, and latency. 3) the beamforming, scheduler, and mobility management should be jointly optimised in LEO satellite communications.

8.2.4 Cross-Layer Design

In satellite networks, the E2E delay consists of multiple delay components in the physical layer, data link layer, and network layer. Existing approaches to communication system design divide the systems into multiple layers according to the Open Systems Interconnec- tion model [185]. These approaches cannot reflect the interactions across different layers, and hence leads to suboptimal solutions. To minimize the E2E delay of the network, we need to design the whole system from a cross-layer approach.

Recently, unsupervised deep learning was applied to solve complicated cross-layer optimisation problems in terrestrial networks [183, 186]. By unsupervised deep learning, a mapping from dynamic environment parameters in wireless networks to the optimal radio resource allocation can be approximated by a DNN. Once the environment parameters (e.g., channel state information and queueing state information) change, the system can obtain the optimal resource allocation from the output of the DNN.

8.2.5 Digital Twins of Satellite Networks

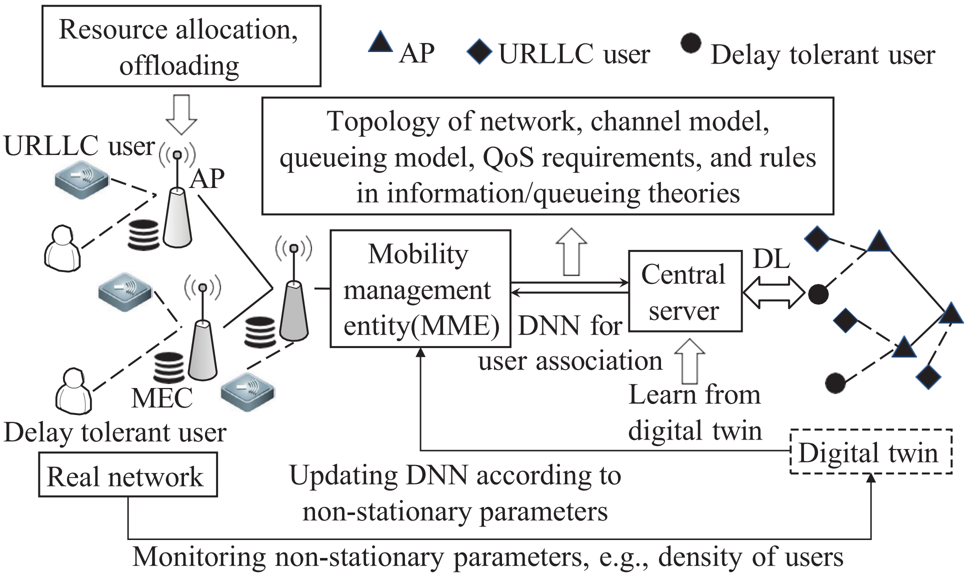

Resource allocation, offloading
▲ AP
◆ URLLC user
● Delay tolerant user
Topology of network, channel model, queueing model, QoS requirements, and rules in information/queueing theories
URLLC user
AP
Mobility management entity(MME)
Central server
DL
DNN for user association
Learn from digital twin
MEC
Delay tolerant user
Real network
Updating DNN according to non-stationary parameters
Digital twin
Monitoring non-stationary parameters, e.g., density of users

Figure 24: Digital twin of a wireless network [187].

The digital twin of a wireless network can serve as a bridge between the model-based analysis and the data-driven deep learning [187]. According to the definition in [188], “a Digital Twin is an integrated multiphysics, multiscale, probabilistic simulation of an as- built vehicle or system that uses the best available physical models, sensor updates, fleet history, etc., to mirror the life of its corresponding real system”. As illustrated in Fig. 24, a digital twin of a terrestrial wireless network exploits the topology, channel and queue- ing models, QoS requirements, and fundamental principles in information and queueing theories to build the simulation platform. Then, the central server keep monitoring non-stationary parameters and update the DNN accordingly. In this way, the control plane can adjust its policies in dynamic wireless networks.

Such a concept can be used to establish a simulation platform of programmable satellite networks, where GEO satellites and LEO satellites serve as the control plane and the data plane, respectively. With LEO constellations, such as SpaceX, OneWeb, and TeleSAT [48], the altitudes range from 1000 to 1325 km. Thus, it is possible to achieve a 10 ms E2E delay from terrestrial users to ground stations that are simultaneously connected to one satellite. For the users and the ground stations served by different satellites, the latencies of the optical ISLs and that of the terrestrial-satellite links should be taken into account. With the help of digital twin, it is possible to optimise routing schemes, user associations, scheduling, and resource allocation subject the QoS requirements in a cross-layer manner by using both optimisation and learning methods.

Compared with terrestrial networks, the control plane latency in satellite networks is much longer. Thus, the system parameters observed by the central server could be outdated. The mismatch between the digital twin and the real-world satellite network may lead to significant performance losses in terms of reliability, latency, AoI, etc. To handle this issue, one potential approach is prediction and communication co-design [189, 190].

8.3 Key Findings and Future Research

The key findings of the literature review in this section are summarized as follows.

- There is no existing software-defined RAN architecture for 5G NTN in the literature.

- Learning-based programmable schedulers can achieve better reliability than heuristic schedulers in real-world terrestrial 5G networks. By developing proper reward and penalty functions, it is possible to meet diverse QoS requirements.

- Although cross-layer design was widely used in terrestrial networks, it has so far not been applied in 5G NTN. One of the reasons is that cross-layer optimisation algorithms are with high computational complexity, and hence can hardly be implemented in real time.

- Deep learning methods allow us to trade online processing delay with off-line computing resources, and can achieve better QoS and resource utilization efficiency than model-based optimisation methods when the models are inaccurate. In addition, GNN-based deep learning applicable in large-scale networks and can be transferred to different typologies, routing, and traffic.

- To optimize routing, scheduling, precoding, and resource allocation, the control plane needs to observe the state information of the satellite network. Due to the long propagation delay, the observed information is outdated. To handle this issue, prediction and communication co-design is a promising approach. Deep learning techniques have yet to be applied to these problems.

Part II
Completed Research Tasks and Results

In this part of the report, we present our completed research tasks and the results ob- tained. In Section 9, we model and simulate spot beam systems and develop adaptive algorithms for spot beam systems. Section 10 presents ths proof-of-concept implementa- tion of combined precoding and user scheduling for the spot-beam forward link. In Sec- tion 11, we present the proof-of-concept implementation of machine learning algorithms to handle satellite channel impairments. In Section 12, the proof-of-concept implemen- tation of FPGA-based fault tolerant SDR for the satellite transponder is presented. In Section 13 we propose an AI-enabled SDN architecture for 5G NTNs, and perform a pre- liminary evaluation of the proposed architecture. We have identified the possible research problems to design and develop the architecture further.

Section 9 focuses on several of the questions raised in Part I concerning adaptive spot beams. In this section, we consider a simple model of non-uniform traffic, and hybrid beamforming (assumed to be done on-board the satellite) to compare the current state- of-the-art Spot Beam technology using multi-horn reflector antennas with that of direct radiating active phased arrays (APAs). For both approaches we consider the rate gains that are achievable from digital multi-beam precoding. While our traffic model is not dynamic, we consider a non-homogeneous spatial distribution of traffic demand and show the benefits that APAs provide from being able to match the (re-configurable) beams to the traffic demands.

Section 10 presents three practical approaches to precoding and user scheduling for the forward link, including proportional fairness precoding which is robust to phase un- certainty, joint precoding and user scheduling algorithms that can adaptively match the real-time traffic demands, and hybrid on-board/on-ground precoding techniques. We also present the concept delay-Doppler domain modulation scheme - OTFS , and the potential gains of applying OTFS to NGSO SatCom systems with fast-moving satellites, through theoretical and numerical analysis.

Section 11 proposes a novel machine learning approach to satellite nonlinear compen- sation. As explained in Part I, Digital pre-distortion (DPD) is a widely used baseband signal processing technique to improve the linearity of the radio transmitter at a high power amplifier (HPA). Before the signal is distorted by the HPA, it will be pre-processed by DPD in order to counteract the nonlinearity of the HPA. HPAs also suffer from mem- ory effects, which are caused by time variations in the amplifier's circuit characteristics. We present a general high power amplifier (HPA) model with memory effect, and propose a neural network (NN)-based digital pre-distortion (DPD) to combat the nonlinearity in HPA.

Section 12 investigates the use of FPGA and artificial intelligence to achieve an SDR implementation for a 5G protocol. By using FPGAs, the computationally intensive tasks including digital signal algorithms, data connections, and fault correction algorithms can be done in parallel using a pipelined approach. The resources and clock cycles can be used more efficiently to improve the overall performance. We provide a simple example concerning 5G NR Cell Search which can be setup as a Simulink model in Matlab and then implemented on a FPGA board. Resulting resource usage is provided in Fig. 51.

Section 13 proposes an artificial-intelligence (AI) enabled SDN architecture for 5G NTN. As observed in Part I, there is no existing software-defined radio network architec- ture for 5G NTN. We segment the 5G NTN into three parts: 1) a 5G NTN core network located in the data centres around the globe, 2) a 5G NTN radio access network including the ground stations, satellites and user equipment, and 3) a 5G NTN satellite networks. To enable programmability of the whole network, we propose to implement agents in net- work devices, where each agent controls the network device and communicates with the AI algorithms running in the data centres. With such an architecture, we can optimise the end-to-end performance across all the networks by designing a joint algorithm controlling the three different parts of 5G NTNs.

9 Modeling and Simulation of Adaptive Spot Beams For a Ka Band Spot-Beam Satellite Over Australia

In this section, we develop adaptive spot-beam techniques for the forward link based on traffic demand. We develop reconfigurable beam hopping and dynamic RF chain allocation algorithms.

9.1 System Model

We consider the coverage area shown in Fig. 25. It consists of 256 squares, each with dimensions 250 km × 250 km. The color of each square represents the number of users in the given square. Each Red square has U_R users, each Yellow square has U_Y users and each Green square has U_G users with $U_G < U_Y < U_R$. In this section, we only consider fixed users. The Grey squares have no users. Fig. 25 is built based on the geographical distribution of population in Australia. It is assumed that the per-user traffic demand is the same for all users. Hence, the traffic demand in each square is characterized by the number of users in the given square. Observe that the coverage area consists of 32 Red squares, 32 Yellow squares and 64 Green squares. The satellite radio frequency (RF) payload has N_D RF chains.

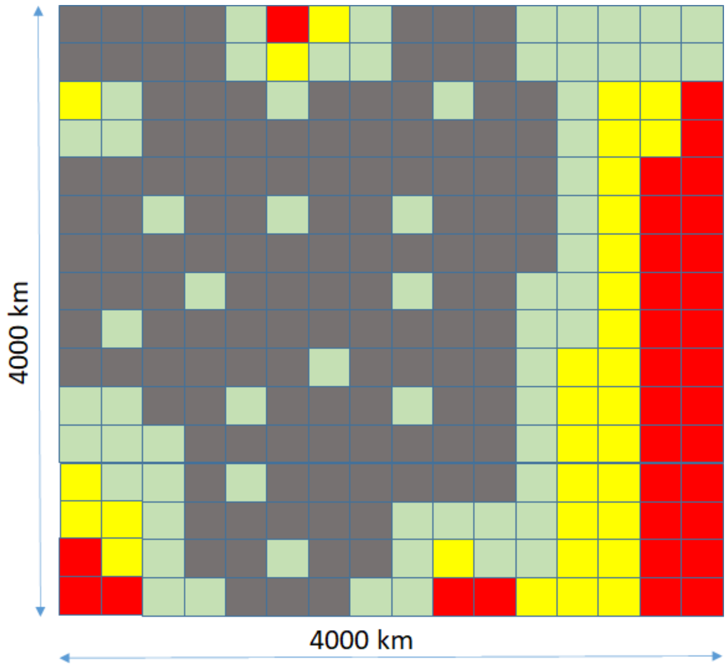


Figure 25: Satellite beams with 4-beam frequency reuse

We consider the following type of user distribution. Each square is divided into M small squares and each small square has $1/M$ of the users of the given square with the users concentrated at the centers of the small squares. Fig. 26 shows the the population distribution in a single square for $M = 4$.

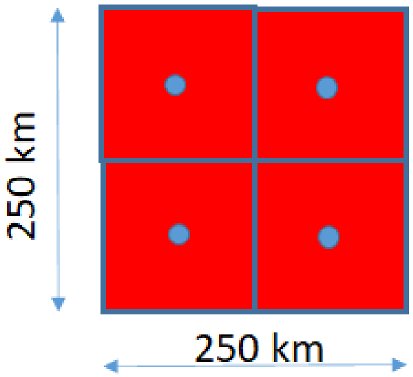


Figure 26: Population distribution in a square for Distribution Type 1 with $M = 4$

9.1.1 Phased-Array Antenna at the Satellite

Current satellite systems use feed-horn antennas with reflectors to provide service to the users. Feed-horn antennas lack flexibility since the beam directions cannot be changed after the satellite is deployed. Therefore, it is difficult to cater for long-term changes in traffic demand (e.g., establishment of new mining towns) efficiently, nor for changes that occur over faster time-scales (e.g. over the course of a day, or even much shorter periods). As an alternative, we use a direct-radiating uniform phased-array antenna at the satellite. A uniform phased array is an array of identical antenna elements with identical magnitude and identical phase shift between adjacent antenna elements. Compared to conventional feed-horn antennas, phased array antennas significantly improve the coverage flexibility due to the fact that the beam directions can be varied on-board dynamically by changing the phase between antenna elements.

Specifically, we use a square uniform planar array (UPA) similar to Fig. 27 [191] with $N_A \times N_A$ isotropic antenna elements at the satellite. It is assumed that the distance between two adjacent antenna elements is $d_x = d_y = \frac{\lambda}{2}$, where λ is the wavelength. If the satellite is in Geostationary orbit, the half-power beam width (HPBW) corresponds to a side of each small square being 0.4 degrees (on the Equator). It can be shown the value of N_A which correspond to HPBW of 0.4 degrees is 252 [191], i.e., the UPA consists of 252 × 252 antennas.

9.1.2 Channel Model with UPAs

The dimensions of the normalized channel matrix (without the free-space loss) $\mathbf{H}(t)$ are $N_D \times N_A^2$, where row i of $\mathbf{H}(t)$ represents the channel gain from all the antennas to user i served at time t. Let $[\mathbf{H}(t)]_{i,(m-1)N_A+n}$ be the element in the i-th row and the $(m-1)N_A+n$-th column of $\mathbf{H}(t)$, which is the channel coefficient between the antenna in the row m and the column n of UPA to the user i served at time t, which is given by

$$[\mathbf{H}(t)]_{i,(m-1)N_A+n} = \exp\left(j(m-1)\frac{2\pi}{\lambda}d\sin\theta_i\cos\varphi_i\right)\exp\left(j(n-1)\frac{2\pi}{\lambda}d\sin\theta_i\sin\varphi_i\right),$$
$$m,n \in \{1,\ldots,N_A\},\ i \in \{1,\ldots,N_D\} \quad (4)$$

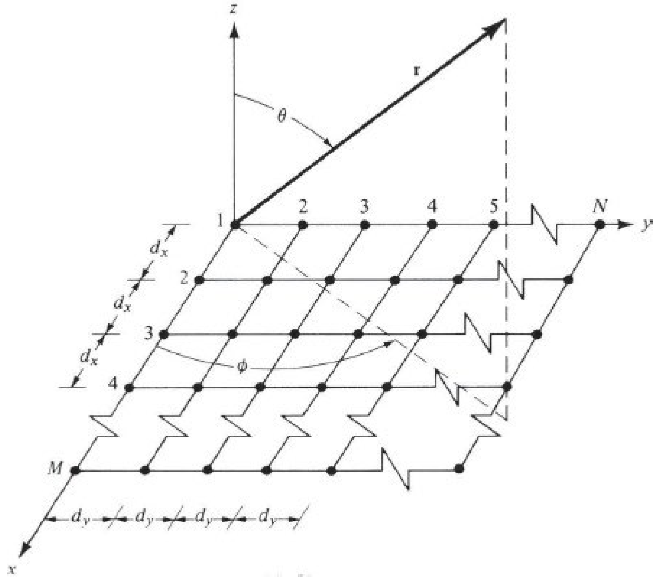


Figure 27: A uniform planar array [191]

where $j = \sqrt{1}$, and θ_i and φ_i are the elevation and azimuth angles for user i served at time t by RF chain i with respect to the UPA, respectively.

Table 8: Antenna parameters of UPAs with different beam diameters.

Beam diameter (km)	Beam width (deg.)	Number of antenna elements	Array Dimensions	Antenna Gain (dB)
1000	1.6	63 × 63	42 cm × 42 cm	40
500	0.8	126 × 126	84 cm × 84 cm	46
250	0.4	252 × 252	1.68 m × 1.68 m	52
125	0.2	504 × 504	3.36 m × 3.36 m	58
62.5	0.1	1008 × 1008	6.72 m × 6.72 m	64

Table 8 shows the antenna parameters of UPAs with different beam diameters. The frequency band and the antenna efficiency are assumed to be 20 GHz and 80%, respectively. It is desirable to have very narrow beams, which have higher gains and lower interference to the users who are not serviced by the given beam. However, narrower beams comes at the cost of larger antenna arrays with elements in the order of millions. On the other hand, wider beams which are more desirable due to smaller array dimen- sions and lower number of antenna elements, cause a higher interference to the users who are not serviced by the beam. To suppress interference caused by wider beams, we can use digital baseband precoding/beamforming. Hence, we consider a hybrid analog-digital beamformer, where the analog beamformer points the beam to the desired user location and the digital beamformer suppresses the interference received from beams destined to other users.

9.2 Hybrid Beamforming Scheme

Fig. 28 shows our hybrid beamforming architecture [192]. It is assumed that each RF chain is connected to all the antennas in the UPA. It consists of a baseband digital precoder matrix $\mathbf{D}(t)$ and a radio-frequency analog beamformer matrix $\mathbf{A}(t)$. The analog beamformer applies a phase shift between two adjacent antenna elements such that the beam is pointed to the desired direction. In Fig. 28, $[\mathbf{A}(t)]_n$ is the n-th column of matrix $\mathbf{A}(t)$, which corresponds to the phase shifts applied to direct the beam of RF chain n. The digital precoder cancels co-channel interference which results from the reuse of the available frequency band by all the RF chains.

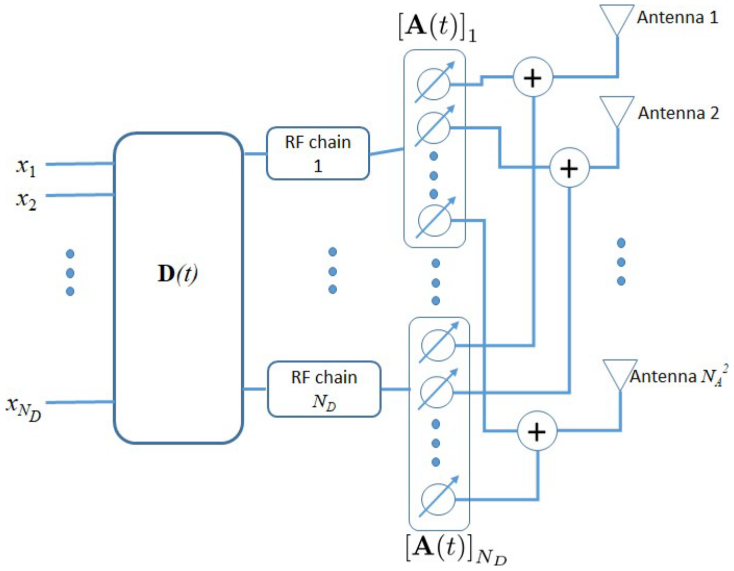


Figure 28: Hybrid beamforming architechture

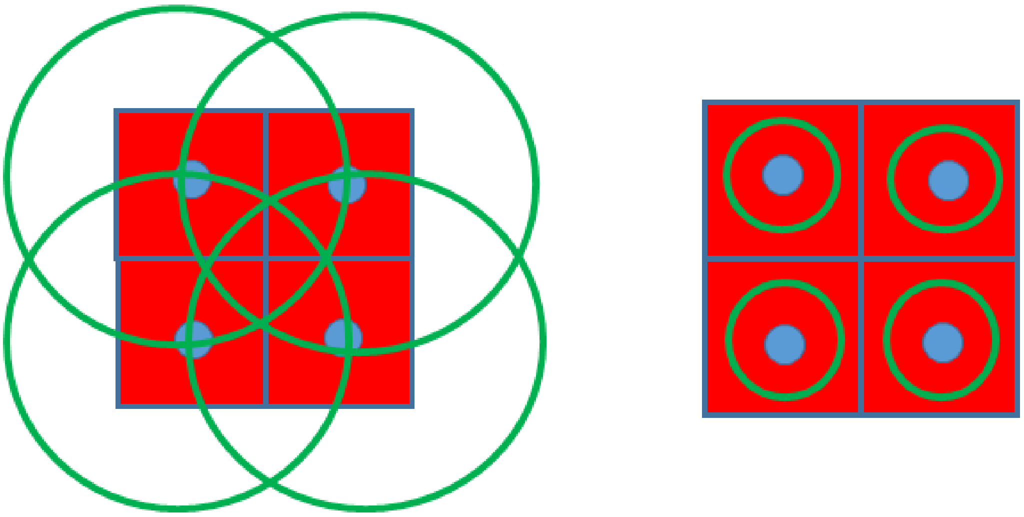
With N_D RF chains, the satellite can serve N_D users at N_D small squares/clusters at a given time. Multiple users in a given small square/cluster are served using time-division multiple access (TDMA). The received signal vector for N_D users served at time t can be given as

$$\mathbf{y}(t) = \sqrt{G_T G_R \Gamma}\,\mathbf{H}(t)\mathbf{A}(t)\mathbf{D}(t)\mathbf{x}(t) + \mathbf{n}(t) \tag{5}$$

where G_T and G_R are the transmit and receiver array gains, respectively, Γ is the free-space loss given by $\Gamma = \left(\frac{\lambda}{4\pi d_0}\right)^2$ where d_0 is the propagation distance, which is 35,786 km, $\mathbf{y}(t)$ is the received signal vector with dimensions $N_D \times 1$, $\mathbf{H}(t)$ is the normalized channel matrix with dimensions $N_D \times N_A^2$, $\mathbf{A}(t)$ is the analog beamforming matrix with dimensions $N_A^2 \times N_D$, $\mathbf{D}(t)$ is the baseband digital beamforming matrix with dimensions $N_D \times N_D$, $\mathbf{x}(t)$ is the symbol vector of the N_D users served at time t, and $\mathbf{n}(t)$ is the noise vector
for the N_D users served at time t. The time index t is used since the set of users served by the satellite changes with time.

9.2.1 Analog Beamforming Scheme

The purpose of RF analog beamforming is to direct the beam of each RF chain to its desired user to maximize the power of the desired signal at the user of the given RF chain. The k-th column of the analog beam forming matrix $[\mathbf{A}(t)]_k$ is the vector of phase



(a) (b)

Figure 29: Analog beamforming technique with wide beams (a) narrow beams (b).

shifts applied to direct the beam generated by the RF chain k. To maximize the received power from the beam k at the desired user, the column k of phase matrix matches to the row k of the channel matrix in (4). Hence, the analog beamforming matrix can be expressed as

$$\mathbf{A}(t) = \frac{1}{N_A^2}\mathbf{H}(t)^\dagger \qquad (6)$$

where $(\cdot)^\dagger$ is the conjugate transpose of a matrix and $\frac{1}{N_A^2}$ is the normalization coefficient.

We consider the following analog beamforming approach. The analog beam center coincides with the location of the user served by the given RF chain at a given time (Fig. 29(a)). Hence, the beam center at a given time coincides with the center of the small sub-square served by the given RF chain. In most of our numerical results, the analog beam spills into adjacent squares causing interference, which is then mitigated via the digital precoding. The exception is the narrow beam scenario (Fig. 29(b)) in which case the 3 dB beamwidth remains inside the small sub-square containing the desired user location with much less spillage into adjacent squares. However, compared to the beams in Fig. 29(a), to generate the narrow beams in Fig. 29(b), 16 times more antenna elements are required.

9.2.2 Digital Beamforming Scheme

The purpose of the baseband digital beamformer is to cancel the interference received at users (the RF chain of each user is across the entire frequency band). The effective channel for the baseband digital beamformer is

$$\bar{\mathbf{H}}(t) = \frac{1}{N_A^2}\mathbf{H}(t)\mathbf{H}(t)^\dagger. \qquad (7)$$

The dimensions of $\bar{\mathbf{H}}(t)$ are $N_D \times N_D$ and the element $\left[\bar{\mathbf{H}}(t)\right]_{i,k}$ is the signal received by user i served at time t from the analog beam destined to user k served at time t, which can be given by [191]

$$\left[\bar{\mathbf{H}}(t)\right]_{i,k} = \frac{1}{N_A^2}\sum_{m=1}^{N_A}\exp\left(j(m-1)\psi_x\right)\sum_{n=1}^{N_A}\exp\left(j(n-1)\psi_y\right) = \frac{1}{N_A^2}\exp\left(j\psi_x\frac{N_A}{2}\right)\frac{\sin\left(\frac{N_A\psi_x}{2}\right)}{\sin\left(\frac{\psi_x}{2}\right)}\exp\left(j\psi_y\frac{N_A}{2}\right)\frac{\sin\left(\frac{N_A\psi_y}{2}\right)}{\sin\left(\frac{\psi_y}{2}\right)} \quad (8)$$

where $\psi_x = \frac{2\pi}{\lambda}(\sin\theta_i\cos\varphi_i - \sin\theta_k\cos\varphi_k)$ and $\psi_y = \frac{2\pi}{\lambda}(\sin\theta_i\sin\varphi_i - \sin\theta_k\sin\varphi_k)$. The phase components of (8) can be eliminated by choosing the central antenna element as the reference point instead of the first element [191].

For baseband digital beamforming, we consider a variant of regularized zero forcing (RZF) beamforming. First, the matrix $\tilde{\mathbf{D}}$ is found by

$$\tilde{\mathbf{D}} = \bar{\mathbf{H}}(t)^{\dagger}\left(\bar{\mathbf{H}}(t)\bar{\mathbf{H}}(t)^{\dagger} + \alpha\mathbf{I}\right)^{-1} \quad (9)$$

where α is a regularization parameter, which was defined in [193]. We consider transmit power constraints for each RF chain separately. Hence, the final beamforming matrix for the users served at time t, $\mathbf{D}(t)$ is obtained by

$$[\mathbf{D}(t)]_k = \frac{P_T}{N_D}\frac{\left[\tilde{\mathbf{D}}\right]_k}{\left\|\left[\tilde{\mathbf{D}}\right]_k\right\|}, \; k = 1, \ldots, N_D \quad (10)$$

where $[\mathbf{D}(t)]_k$ is the k-th column of $\mathbf{D}(t)$ and P_T is the total transmit power. Thus, the directions of the columns in $\mathbf{D}(t)$ are preserved and normalized according to the per-RF chain power.

The signal received by the user i is

$$y_i(t) = \sqrt{G_T G_R \Gamma}\left[\bar{\mathbf{H}}(t)\right]^i[\mathbf{D}(t)]_i x_i + \sqrt{G_T G_R \Gamma}\sum_{l=1, l\neq i}^{N_D}\left[\bar{\mathbf{H}}(t)\right]^i[\mathbf{D}(t)]_l x_l + n_i(t) \quad (11)$$

where $\left[\bar{\mathbf{H}}(t)\right]^i$ is the i-th row of $\mathbf{H}(t)$ and x_i is the data symbol for user i. The signal-to-interference-plus-noise ratio (SINR) at user i is

$$\gamma_i(t) = \frac{G_T G_R \Gamma\left|\left[\bar{\mathbf{H}}(t)\right]^i[\mathbf{D}(t)]_i\right|^2}{BN_0 + G_T G_R \Gamma\sum_{l=1, l\neq i}^{N_D}\left|\left[\bar{\mathbf{H}}(t)\right]^i[\mathbf{D}(t)]_l\right|^2} \quad (12)$$

where B is the total user bandwidth and N_0 is the noise power spectral density.

9.3 Reconfigurable Spot Beam Techniques

As we mentioned earlier, in the coverage area given in Fig. 25, the traffic demand varies between squares of different color. Therefore, the allocation of the same amount of spectrum/time resources to each square will result in over-allocation of resources per user in Green squares and insufficient allocation of resources per user in Red squares. To address this issue, we develop reconfigurable spot beam techniques based on traffic demand. We first develop a beam hopping technique, which adapts the amount of time allocated for each square based on the number of users in the given square and the user channel con- ditions. We then develop a dynamic RF chain allocation scheme where the number of RF chains allocated to each color depends on the aggregate traffic demand of the given color. Although we use the terms “adapt” and “dynamic” to describe these schemes, in the simple coverage model shown in Fig. 25, the traffic pattern is fixed, whilst being spatially inhomogeneous. However, it is clear that our techniques can be made adaptive and dynamic, should the traffic pattern change over time.

9.3.1 Reconfigurable Beam Hopping Scheme

In our setup, the traffic demand in each square is characterized by the number of users in the square. Recall that there are 32 Red squares, 32 Yellow squares and 64 squares in the coverage map given in Fig. 25. Hence, there are $32M$ different Red user locations, $32M$ different Yellow user locations and $64M$ different Green user locations, where M is the number of user locations per square.

Under our beam hopping scheme, each RF chain serves one Red user location, one Yellow user location and two Green user locations. Also, it serves one user location at a time, using time hopping to cover the above users in the various squares. For simplicity, it is assumed that all the RF chains first cover the Red squares, then the Yellow squares, and finally the Green squares during a beam hopping cycle. The proposed beam hopping technique determines the dwell time of RF chains with respect to each color.

Let the dwell time of RF chains in Red, Yellow and two Green user locations be τ_R, τ_Y, τ_{G_1} and τ_{G_2}, respectively. Let the beam hopping window be $\tau_R + \tau_Y + \tau_{G_1} + \tau_{G_2} = \tau_S$, where τ_S is constrained by the maximum tolerable delay between two successive coverages for the user application. Since users are concentrated in the centers of small sub-squares, the SINR at a given user location is assumed to be the same for all the users at that location. Let $\gamma_{R,i}$, $\gamma_{Y,i}$, $\gamma_{G_1,i}$ and $\gamma_{G_2,i}$ be the user SINR for small Red, Yellow and Green squares served by the RF chain i, respectively, which are computed as in (20). Recall that U_R, U_Y and U_G are the number of users in each large Red, Yellow and Green square, respectively, which characterizes the traffic demand in each small square since the traffic demand for each user is assumed to be the same.

The objective of our proposed reconfigurable beam hopping algorithm is to optimize the ratio between the offered capacity and the traffic demand [72] for users in different colors through the dwell times of RF chains in each color. The objective function for reconfigurable beam hopping can be given as follows.

$$\begin{aligned}
&\max \quad \min_{\tau_R, \tau_Y, \tau_{G_1}, \tau_{G_2}} \left(\frac{C_R}{U}, \frac{C_Y}{U_Y}, \frac{C_{G_1}}{U}, \frac{C_{G_2}}{U} \right) \\
&\text{s.t.} \quad \sum_{c \in \{R, Y, G_1, G_2\}} \tau_c = \tau_S, \\
&\qquad C_c = \frac{B\tau_c}{\tau_S} \log_2 (1 + \gamma_c), \quad \gamma_c = \min(\gamma_{c,1}, \ldots, \gamma_{c,N_D}) \\
&\qquad \tau_{c,i} > 0 \quad c \in \{R, Y, G_1, G_2\}, i \in \{1, \ldots, N_D\},
\end{aligned} \tag{13}$$

where B is the total bandwidth in the forward link. The problem in (13) can easily be generalized into more than three types of traffic demands.

Observe that in (13) we consider the user location with worst SINR of each color when determining the offered capacity to ensure that at a given time, all the RF chains serve the user locations of a single color, which simplifies computation of the effective channel matrix in (8). However, the users of a given color in locations with better SINR than the worst-SINR user location for the same color will be offered more capacity than the worst-SINR user. Therefore, the per-user offered capacity slightly varies between different user locations of the same color.

9.3.2 Dynamic RF Chain Allocation Scheme

In the reconfigurable beam hopping algorithm described above, at a given time, the users of only one color are served. From Fig. 25, most user locations are located close to other user locations of the same color, which results in high interference. Instead of allocating all the RF chains to a single color at a time, if RF chains are allocated based on traffic demand, the users in moderate and lower traffic-demand squares will experience significantly lower interference than in the beam hopping technique since the number of RF chains allocated to these colors will be significantly lower than the number of user locations. The users in high-traffic squares will also experience a slight reduction in interference since the number of RF chains allocated to them at a time is slightly less than N_D. Based on this premise, we propose a dynamic RF chain allocation algorithm. In this algorithm, each color is allocated a number of RF chains based on the traffic demand of the color, and for a given RF chain allocation, each RF chain serves users of a single color. In this scheme, users of different colors are served simultaneously.

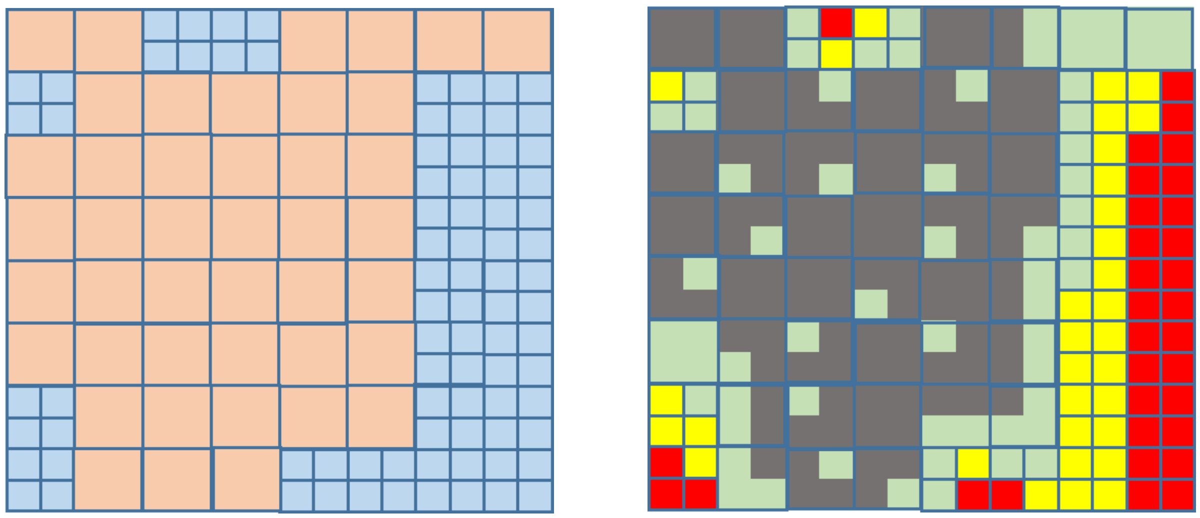
Let N_R, N_Y and N_G be the number of RF chains allocated for Red, Yellow and Green squares, respectively. In the proposed dynamic RF chain allocation algorithm, we optimize the ratio between the total average capacity of each color and the total traffic demand of each color. The optimization problem is given as follows.

$$\begin{aligned} &\max \min_{N_R, N_Y, N_G} \left(\frac{C_R}{U_R S_R}, \frac{C_Y}{U_Y}, \frac{C_G}{U_G S_G} \right) \\ &\text{s.t.} \quad N_R + N_Y + N_G = N_D, \\ &\qquad N_R \in \mathbb{Z}^+, N_Y \in \mathbb{Z}^+, N_G \in \mathbb{Z}^+ \end{aligned} \tag{14}$$

where C_R, C_Y and C_G are the total average rates offered for all Red, Yellow and Green squares, respectively, and S_R, S_Y and S_G are the number of Red, Yellow and Green squares, respectively. Note that C_R, C_Y and C_G are functions of N_R, N_Y and N_G, re- spectively. The problem in (14) is solved through an exhaustive integer search over all possible combinations of N_R,N_Y and N_G which satisfy the constraints in (14).

For a given combination of N_R, N_Y and N_G, the set of user locations to be served in each color at each time slot are selected randomly from all user locations of the given color. More sophisticated user selection algorithms which target to minimize interference should improve the performance further at the cost of higher computational complexity.

(14) can also be generalized to more than three types of traffic demands. However, the computational complexity of exhaustive search increases as the number of traffic types (colors) increases.



(a) (b)

Figure 30: Spot beams used in the fixed beam system (a) and the traffic demand in each fixed beam (b).

9.4 Numerical Results

9.4.1 Simulation Setup

In this section, we present the numerical results obtained for the proposed adaptive spot beam techniques. We compare the performance results of traffic demand based adaptive resource allocation techniques which use phased array antennas with the performance of fixed-beam systems which allocate the same amount of resources to all the squares. The simulation parameters are given in Table 9.

Table 9: Simulation parameters used to test the performance of proposed adaptive re- source allocation techniques for spot-beam forward link.

Parameter	Value
Number of active users in a Red square (U_R)	1000
Number of active users in a Yellow square (U_Y)	100
Number of active users in a Green square (U_G)	10
Satellite Orbit	GEO (35,786 km)
Frequency band	Ka (20 GHz)
Forward link bandwidth allocated to the satellite (B)	500 GHz
Receiver antenna gain (G_R)	42 dBi
Transmit power (P_T)	10 kW
Free space loss (Γ)	213 dB
Distance between two adjacent antenna elements (d)	$\frac{\lambda}{2}$ = 0.667 cm
Receiver noise temperature	290 K

For the fixed-beam system, we use the beams shown in Fig. 30(a), which is based on the spot beams used by Sky Mesh to provide internet access to its customers in Australia, as shown in Fig. 1 [10]. In Fig. 30(a), the areas with high population density are covered by beams with diameter 250 km and the areas with low population density are covered by beams with diameter 500 km. Fig. 30(b) shows the traffic demand of each fixed beam by overlapping Fig. 30(a) on Fig. 25. There are 127 beams in Fig. 30(a). Observe that there are some wide beams with zero traffic demand in the fixed-beam system. It is assumed that zero power is allocated by the satellite to these beams in the fixed-beam system. There are 110 beams with non-zero traffic demand in Fig. 30(a).

We implement the following spot-beam technique in our simulations to compare the performance.

1. Reflector-antenna fixed beams with 28 RF chains- This technique uses fixed-beams generated by feed-horn reflector antennas (Fig. 30(a)) with four-cell frequency reuse. Recall that no power is allocated for beams with zero traffic demand. With four- cell reuse, 28 RF chains are used to serve the remaining 110 beams with non-zero traffic demand. With four-cell reuse in feed-horn antennas, it can be shown that the system is noise limited. Each beam is allocated the same amount of frequency and time resources.

2. Reflector-antenna fixed beams with 110 RF chains - In this technique, fixed-beams generated by feed-horn reflector antennas are considered with each beam with traffic demand in Fig. 30(a) is served with the same frequency band (universal frequency reuse). In this case, 110 RF chains are used to serve all the squares simultaneously. RZF precoding is used to suppress interference.

3. Reconfigurable beams with 32 RF chains and beam hopping- This technique uses reconfigurable beams generated by phased arrays with the proposed reconfigurable beam hopping algorithm. Similar to the first technique, four-color frequency reuse is used, which requires 32 RF chains to be used at the spot-beam transmitter.

4. Reconfigurable beams with 128 RF chains and beam hopping - In this and the two techniques that follow, universal frequency reuse across beams is used to improve the spectral efficiency.

5. Reconfigurable beams with 128 RF chains and RF chain allocation - In this tech- nique, we implement the proposed dynamic RF chain allocation algorithm with 128 RF chains.

6. Reconfigurable narrow beams with 128 RF chains and beam hopping- Compared to the previous technique which consider a beam diameter of 250 km, this technique uses beams with diameter 62.5 km. Hence, this technique requires 16 times more antenna elements than the previous systems. 128 RF chains are used at the spot- beam transmitter.

9.4.2 Performance Criterion

Let τ be a per-user data rate threshold, which is the rate guaranteed by the communication service provider to its users. We define the function $F_c(\tau)$ as follows for a given spot-beam technique out of the six spot-beam techniques listed earlier in this section.

$$F_c(\tau) = \frac{1}{S_c} \sum_{k=1}^{S_c} (C_{c,k} < \tau \quad , \quad c \in \{\text{Red,Yellow,Green}\} \qquad (15)$$

where S_c is the number of squares with color c, $C_{c,k}$ is the rate offered by the given spot- beam technique for the square k of color c, U_c is the number of users in a square of color

c and $_{\text{(condition)}}$ is the indicator function, which is equal to one if the condition is satisfied and zero otherwise. Since the number of users is the same for all the squares of a given color, $F_c(\tau)$ is the fraction of users of color c for which the achievable rate provided by the spot-beam technique considered is less than or equal to the target rate. The users in squares where target rate τ is not provided by the given spot-beam technique experience a rate lower than the target rate. $F_c(\tau)$ approaches 1 as τ increases.

The function $F_c(\tau)$ enables the satellite service providers to choose the most suitable spot-beam technique based on the availability and cost of resources (e.g., the number of RF chains and the number of antenna elements) to satisfy a given rate requirement and traffic demand. Figs. 31 shows the fraction of all users for which the achievable rate is less than or equal to the target rate threshold τ for the six spot-beam techniques considered. Figs. 32, 33 and 34 show the corresponding fractions for the users in Red, Yellow and Green squares, respectively.

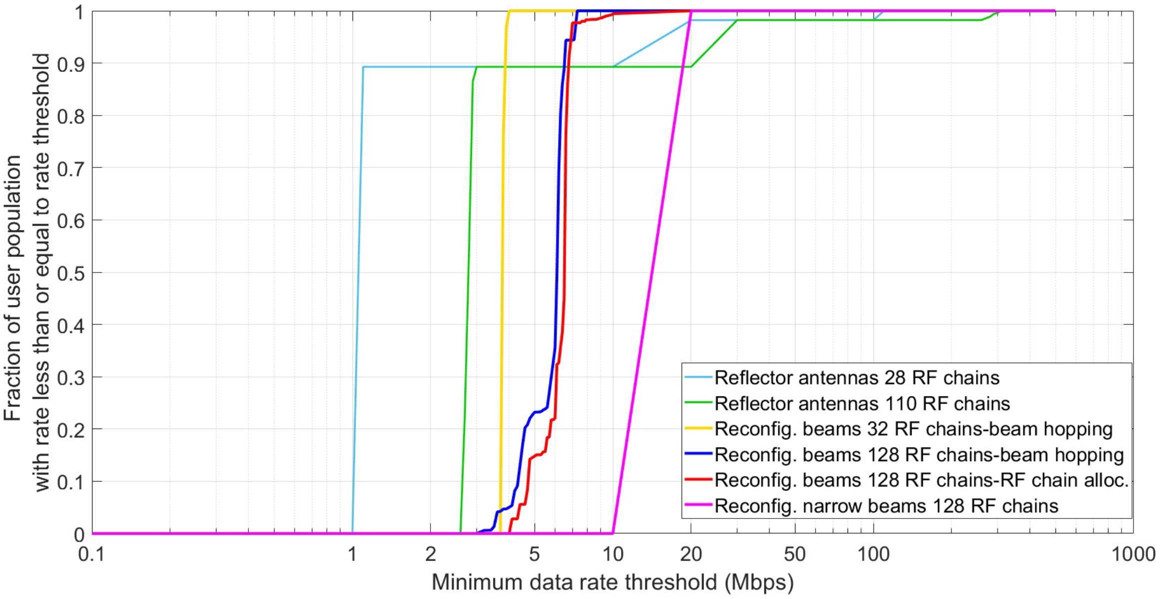


Figure 31: The fraction of all users with achievable rate less than or equal to the rate threshold τ for the six spot-beam techniques considered.

9.4.3 Discussion

The main insights drawn from the simulation results in Figs. 31-34 can be listed as follows.

- From Fig. 31, for reconfigurable spot-beam techniques, the curve rises rapidly from zero to one within a small range of per-user data rate threshold. This indicates that traffic-demand based reconfigurable spot-beam techniques provide similar per-user data rates across the three traffic demand types considered. On the other hand, fixed-beam techniques rises step-by-step over a large range of rate threshold since they deliver very high per-user rates for the Green-square users and low per-user rates for the Red-square users, as shown in Fig. 32 and 34. By definition, fixed-beam techniques allocate the same amount of resources for all the squares, which favour the users in the areas with lower traffic demand.

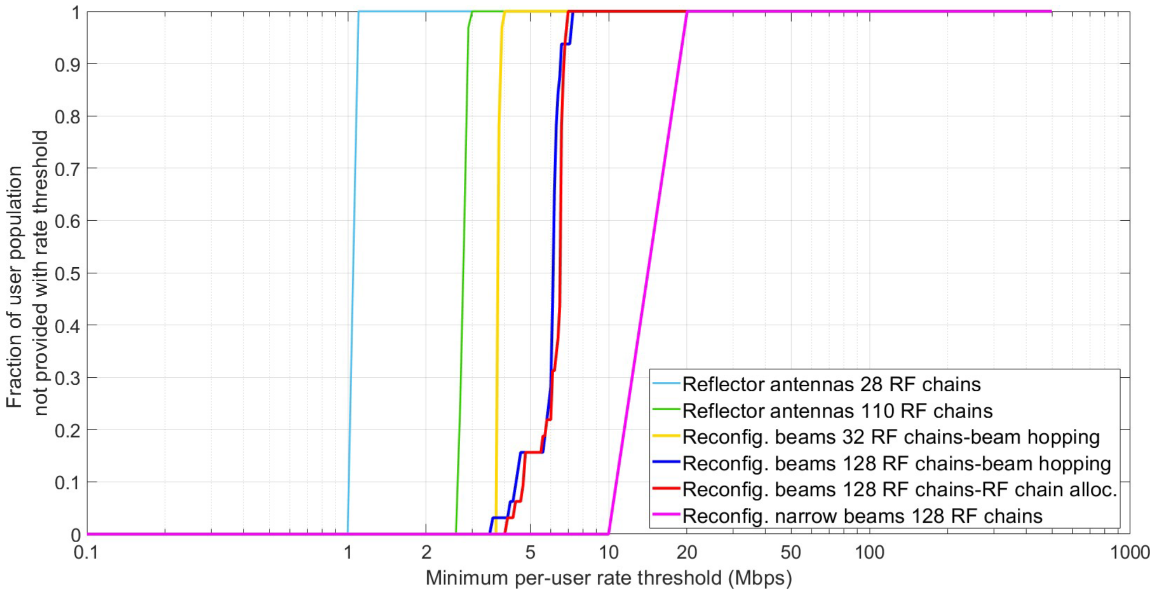


Figure 32: The fraction of users in Red squares with achievable rate less than or equal to the rate threshold τ for the six spot-beam techniques considered.

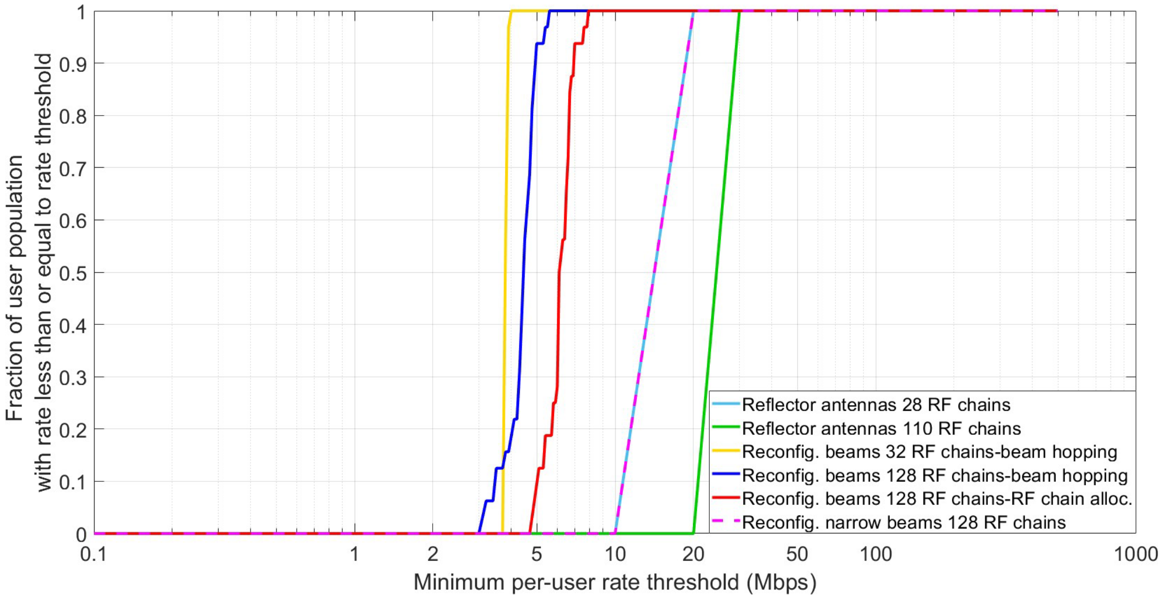


Figure 33: The fraction of users in Yellow squares with achievable rate less than or equal to the rate threshold τ for the six spot-beam techniques considered.

- For users in Red squares which has the highest traffic demand, reconfigurable spot- beam techniques with only 32 RF chains offer a better per-user rate than fixed beams with 128 RF chains, which shows that adaptive techniques are more adept at allocating limited resources more fairly to all the users.
- The systems which use 128 (or 110) RF chains outperforms systems which use 32 (or

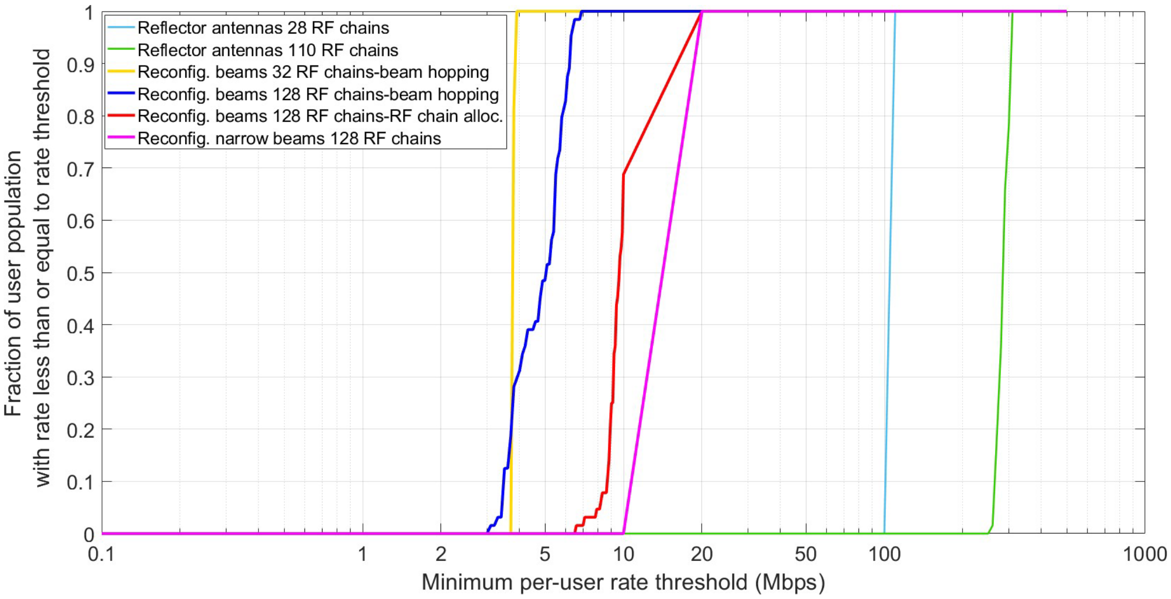


Figure 34: The fraction of users in Green squares with achievable rate less than or equal to the rate threshold τ for the six spot-beam techniques considered.

28) RF chains with similar settings. Note that this is not straight forward since more RF chains means significantly higher interference due to higher reuse of frequency. Furthermore, per-RF chain power is reduced due to total power constraint. Digital precoding enables the suppression of interferfence between beams arising from the higher reuse of frequency enabling the schemes using higher numbers of RF chains to achieve higher per-user rates than with the smaller number of RF chains.

- From Figs. 31-34, the proposed dynamic RF chain allocation algorithm outperforms the reconfigurable beam hopping algorithm in all traffic demand types. In the reconfigurable beam hopping algorithm, at a given time, all RF chains serve a single color, where as in dynamic RF chain allocation, the number of RF chains which serve a given color depends on the traffic demand. Therefore, the interference in dynamic RF chain allocation is lower compared to the reconfigurable beam hopping algorithm. This effect is more prominent in the Yellow and Green squares, where the number of RF chains allocated are significantly lower than the number of user locations due to lower traffic demand.

- In terms of delivering high per-user data rates in all squares, Reconfigurable narrow beams with 128 RF chains and beam hopping is superior to all the other techniques. The narrow beams reduce interference to the other users significantly as shown in Fig. 29(b), and the remaining interference can be suppressed through dig- ital precoding. This provides a gain in capacity of about 3 times over the wider
beam approach. However, this technique requires a phased array with 1008 × 1008 antennas, as compared to the wider beams requiring an array of 252×252 antennas.

- Using reconfigurable APAs increases rates in congested squares by a factor of about 4, as compared to using fixed beams. Digital precoding with both fixed and recon- figurable systems can increase capacity by a factor of about 2.

9.5 Conclusions and Future Work

In this section, we developed traffic-demand based adaptive spot-beam techniques for the forward link and simulated their performance in a coverage area which is built based on the geographical distribution of population in Australia. Three levels of traffic demand were considered for the coverage area based on user density, which are high, moderate and low, represented by Red, Yellow and Green squares, respectively.

We considered a reconfigurable direct-radiating phased-array antenna at the satellite to improve coverage flexibility. We used hybrid analog-digital beamforming to direct beams to the desired users and to suppress co-channel interference resulting from universal frequency reuse. Two traffic-demand based adaptive spot beam techniques were proposed, a reconfigurable beam hopping algorithm and a dynamic RF chain allocation algorithm.

Key insights of this chapter can be summarized as follows.

- The main advantage of our proposed adaptive spot beam systems over the conven- tional reflector antenna systems is dynamic reconfigurability of the system based on the variations of traffic demand which could occur over a larger time scale (e.g., establishment of a new mining town) or over much smaller time scales, even down to mm-sec depending on channel state information (CSI) acquisition and beam switching speeds. Our proposed adaptive spot beam systems can allocate resources efficiently to cater these slow and fast variations in traffic demand.

- Our system uses a reconfigurable direct-radiating phased array antenna, where the beam direction can be configured dynamically by changing the phase shifts between the currents feeding the adjacent antenna elements. It is an efficient method to serve a moving vehicle with a very high traffic demand (e.g., a passenger aircraft, a cruise ship). By changing the input phase, we can always point the beam center to the moving vehicle.

- Our proposed adaptive spot beam techniques deliver similar per-user rates for the users in different traffic-demand areas. However, conventional fixed-beam systems allocate the same amount of resources for all the squares, which results in signif- icantly higher per-user rates for the users in lower traffic-demand areas and lower per-user rates in higher traffic-demand areas.

- The use of digital precoding along with analog beamforming (hybrid beamforming) is essential to suppress high interference resulting from large numbers of RF chains and relatively wide beams. Consequently, the systems which use 128 RF chains significantly outperform systems which use 32 RF chains with similar set- tings. Note that this is not straight-forward since the systems with 128 RF chains has a lower per-RF chain power (due to total on-board transmit power constraint) and significantly higher interference.

- Out of our two proposed adaptive spot-beam techniques, the dynamic RF chain allocation algorithm outperforms the reconfigurable beam hopping algorithm in terms of per-user rates, since interference present in the RF chain allocation algorithm is lower. This effect is more prominent in Yellow and Green squares where the number of allocated RF chains is significantly lower than the number of user locations.

- Our proposed adaptive spot beam system with narrow beams with beam diameter 62.5 km and 128 RF chains provides the best per-user rates for all users in the coverage area. This is due to increased antenna gains and reduced interference caused by narrow beams. However, it requires 1008 × 1008 antenna elements in the phased array. Nevertheless, we believe that antenna arrays with the number of elements in the order of millions should not be ignored due the potential benefits of narrower beams and the advancements in the monolithic microwave integrated circuit (MMIC) technology.

Future research directions emerge from our work can be listed as follows.

- So far we have only considered three types of traffic demands based on the user density in each area. Moreover, we assumed that the traffic demand of a given ge- ographical location remains the same throughout. Furthermore, we have assumed homogeneous per-user traffic demand. In practical systems, the traffic demand of a given location varies with time based on user activity. It is important to explore the time scale on which resources should be adapted to the underlying traffic demands. In practice, user traffic demands are heterogeneous, which vary based on user appli- cations. As an example, users involved in video streaming demand higher data rates than users who browse web at a given time. Also, the geographical distribution of users in Australia is more complicated than the simple model considered in Fig. 26. Practical systems designed to Australia should use actual user distributions in each individual square. It is vital to incorporate these intricacies into our traffic model when developing adaptive spot beam techniques for practical applications.

- From Figs. 31-34, the proposed dynamic RF chain allocation algorithm outperforms the reconfigurable beam hopping algorithm in all traffic demand types, despite the fact that the current dynamic RF chain allocation algorithm selects users to be served in each color in each time slot randomly. Clearly, random user selection is sup-optimal since we do not consider inter-user interference. A better approach is to schedule users based on minimizing the inter-user interference. Dynamic RF chain allocation with adaptive user scheduling based on minimizing interference needs to be investigated in detail in the future to identify potential performance gains.

- So far, we have assumed that the Grey squares in Fig. 25 have zero traffic demand since we have only considered the fixed users. In practice however, the traffic demand in dark areas is not zero as there can be a small number of moving users in these areas. It is vital to provide coverage to them since these areas are usually not covered by cellular operators. To provide coverage to these users using our proposed techniques, some methods are required to be developed to estimate the presence of users and their traffic demand in the Grey squares.

10 Proof of Concept Implementation of Precoding and User Scheduling for the Forward Link and Or- thogonal Time-Frequency Space Modulation (OTFS) for NGSO SatComs

In this section, we present three practical topics on precoding and user scheduling for the forward link, including proportional fairness precoding which is robust to phase un- certainty, joint precoding and user scheduling algorithms that can adaptively match the real-time traffic demands, and hybrid on-board/on-ground precoding techniques. We also present the concept delay-Doppler domain modulation scheme - OTFS , and the potential gains of applying OTFS to NGSO SatCom systems with fast-moving satellites, through theoretical and numerical analysis.

10.1 Robust Proportional Fairness Precoding

As discussed in Section 2.3, in practice, due to the characteristics of satellite systems, the desirable perfect CSI is often hard to obtain, which can bring significant performance degradation of precoding and user scheduling. Due to the long-distance LoS propagation, users served by one beam generally have similar amplitudes of channel vectors, but dif- ferent phases. Therefore, in this section, we propose a robust precoding design aiming to mitigate the performance degradation caused by phase mismatches between estimated and actual CSI.

In this work, we formulate a outage constrained proportional fairness precoding prob- lem, which is robust to phase uncertainty. The proportional fairness power control tech- nique [62, 194] guarantees the fair resource allocation among users, and hence is resource- saving. We have successfully transformed the optimization problem to convex, and pro- posed an algorithm with low computational complexity.

10.1.1 System Model

Consider a geosynchronous earth orbit (GEO) satellite of which the coverage area is near the equator. The number of feeder antennas is N_t and the number of the user groups is N_t. Assuming that the number of users in each group is U_k, one user can belong to
only one user group. Let $\mathsf{G} = [\mathsf{G}_1, ..., \mathsf{G}_{N_t}]$ stand for the index set of scheduled users, where G_k represents the set of scheduled user index of the $k\text{-}th$ user group, $k \in [1, ..., N_t]$. Thus, card($\mathsf{G}_k \leq U_k$). The symbols to be transmitted are given by a complex vector $\mathbf{s} \in \mathbb{C}^{N_t \times 1}$, where $\mathbf{s} \sim \mathcal{CN}(0, \mathbf{I}_N)$ has the unit power. The precoded signal is represented by $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$, the value of $\mathbf{x}$ is given by [25]

$$\mathbf{x} = \mathbf{W}\mathbf{s} \tag{16}$$

where $\mathbf{W}$ fj. $[\mathbf{w}_1, ..., \mathbf{w}_{N_t}] \in \mathbb{C}^{N_t \times N_t}$ is the precoding matrix, another way to write the precoded signal can be

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{w}_k s_k \tag{17}$$

where $\mathbf{s} = [s_1, ..., s_{N_t}]$. The received signal of users is represented by $\mathbf{y}$, which is given by [25]

$$\mathbf{y} = \mathbf{H}^H\mathbf{x} + \mathbf{n} \tag{18}$$

where $\mathbf{H}$ denotes the channel matrix. For convenience, $\mathbf{H}_k \in \mathbb{C}^{N_t\times U_k}$ is defined as the channel matrix of the k-th user cluster. Thus, $\mathbf{H} = [\mathbf{H}_1, ..., \mathbf{H}_N]$. Let $\mathbf{h}_{k,q} \in \mathbb{C}^{N_t\times 1}$ stands for the channel vector of the q-th user in the k-th user group, $q \in [1, ..., U_k]$, the channel matrix of the k-th user cluster can be written as $\mathbf{H}_k = [\mathbf{h}_{k,1}, ..., \mathbf{h}_{k,U_k}]$. The received signal of the q-th user in the k-th user group is further expressed by [25]

$$y_{k,q} = \mathbf{h}^\dagger_{k,q}\boldsymbol{x} + N_0 = \mathbf{h}^\dagger_{k,q}\sum_{l=1}^{K}\mathbf{w}_l s_l + N_0 = \mathbf{h}^\dagger_{k,q}\mathbf{w}_k s_k + \mathbf{h}^\dagger_{k,q}\sum_{l=k}^{K}\mathbf{w}_l s_l + N_0 \qquad (19)$$

with the expression in the equation (19), the SINR of each user is given by [25]

$$\Gamma_{k,q} = \frac{|\mathbf{h}^\dagger_{k,q}\mathbf{w}_k|^2}{\sum_{l=k}^{K}|\mathbf{h}^\dagger_{k,q}\mathbf{w}_l|^2 + N_0} \qquad (20)$$

the onboard power of the n-th feeder antenna is $[\sum_{k=1}^{K}\mathbf{w}_k\mathbf{w}^\dagger_k]_{n,n}$. Let $\hat{\mathbf{h}}$ represents the estimated channel state information at the transmitter (CSIT), assuming there is only phase variation between $\hat{\mathbf{h}}$ and $\mathbf{h}$, the relation between the phase of the estimated and real channel of the q-th user in the k-th user cluster is given by $\Delta\boldsymbol{\theta}_{k,q} \sim \mathcal{N}(0, \sigma^2_{k,q}\boldsymbol{I}_{N_t})$. $\sigma^2_{k,q}$ stands for the variance of the phase error. Thus, the real channel state information (CSI) is given by [195]

$$\mathbf{h}_{k,q} = \hat{\mathbf{h}}_{k,q}\ 0\ \mathbf{e}_{k,q} = \mathrm{diag}(\hat{\mathbf{h}}_{k,q})\mathbf{e}_{k,q} \qquad (21)$$

where $\boldsymbol{e}_{k,q}$ fj. $\exp\{j\Delta\boldsymbol{\theta}_{k,q}\}$, define $\mathbf{E}_{k,q}$ fj. $\mathbf{e}_{k,q}\mathbf{e}^\dagger_{k,q}$ the long term correlation matrix of $\mathbf{e}_{k,q}$ is given by $\mathbf{A}_{k,q} = \mathrm{E}(\mathbf{E}_{k,q})$, where $\mathrm{E}(\cdot)$ denote the expectation. $\mathbf{A}_{k,q}$ is expressed as [195]

$$[\mathbf{A}_{k,q}]_{mn} = \begin{cases} 1, & m = n, \\ \exp\{-\sigma^2_{k,q}\}, & m \neq \end{cases} \qquad (22)$$

The equation (22) is valid when the value of $\sigma_{k,q}$ is small.

10.1.2 Problem Formulation

Assuming that the number of scheduled users in each user group is Q where $Q \leq U_k$, $k \in [1, ..., N_t]$, the outage constrained weighted proportional fairness problem is formulated as

$$\mathrm{P}: \max_{\mathbf{r},\mathbf{P}_{\text{cluster}},\{\mathbf{v}\}_{k=1}^{N_t}} \log_2\Big(\sum_{k=1}^{N_t} r_k\gamma_k\Big),$$

$$s.t.\ \mathrm{C1}: \Pr\Big\{\frac{\Gamma_{k,q}}{\gamma_k} \geq r_k\Big\} \geq 1 - p_{\text{outage}}, \forall k \in [1, ..., N_t], \forall q \in [1, ..., Q],$$

$$\mathrm{C2}: \Big[\sum_{k=1}^{K}\mathrm{P}_{\text{cluster},k}\mathbf{v}_k\mathbf{v}^\dagger_k\Big]_{nn} \leq \mathrm{P}_{\text{antenna,n}}, \forall n \in [1, ..., N_t], \qquad (23)$$

where $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_N] \in \mathbb{R}^{N_t\times 1}$ denotes to the SINR weight for each user group. The constraint C_1 is the outage probability constraint, $\mathbf{r} = [r_1, ..., r_{N_t}]$ is the ratio between the outage threshold and required SINR, $\{\mathbf{v}\}_{k=1}^{N_t}$ is the set of normalized precoding matrix.

$\Gamma_{k,q}$ denotes the signal-to-noise-and-interference-ratio of the q-th users in the k-th beam given in equation (20), it can be rewritten by

$$\Gamma_{k,q} = \frac{P_{\text{cluster},k}\mathbf{h}^{\dagger}_{k,q}\mathbf{v}_k\mathbf{v}^{\dagger}_k\mathbf{h}_{k,q}}{\sum_{l\neq k}^{N_t} P_{\text{cluster},l}\mathbf{h}^{\dagger}_{k,q}\mathbf{v}_l\mathbf{v}^{\dagger}_l\mathbf{h}_{k,q} + N_0}, \tag{24}$$

where $\mathbf{P}_{cluster} = [P_{\text{cluster},1}, ..., P_{\text{cluster},N_t}]$ denotes the power allocated for each user cluster. The relationship between $\mathbf{v}_k$ and $\mathbf{w}_k$ in equation (20) is given by

$$\mathbf{w}_k = \sqrt{P_{\text{cluster},k}}\mathbf{v}_k \quad \forall k \in [1, ..., N_t] \tag{25}$$

The constraint C_2 is the per-antenna power constraint, $\mathbf{P}_{\text{antenna}} = [\mathbf{P}_{\text{antenna},1}, ..., \mathbf{P}_{\text{antenna},N_t}]$ denotes the maximum onboard power for the N_t feeder antennas. The per-antenna power constraint is selected herein, as individual amplifiers for each antenna which prevents power-sharing is adopted. This scheme has a low cost compared to the flexible amplifier which can enable power sharing among antennas [25]. The problem P has the following input constants: $\mathbf{P}_{\text{antenna}}$, and $\boldsymbol{\gamma}$. The problem P constraint can be converted to problem P_r by transforming the constraint C_1 to a convex form [35, 36]

$$\begin{aligned}
P_r: \; & \max_{\mathbf{r},\mathbf{P}_{\text{cluster}},\{\mathbf{v}\}_{k=1}^{N_t}} \log_2\Big(\sum_{k=1}^{N_t} r_k\gamma_k\Big), \\
s.t.\; & C_1: \|\mathbf{G}_{k,q}^{\frac{1}{2}}\mathrm{vec}(\mathbf{C}_k^H)\|_2 \le \frac{1}{b_{k,q}}\Big(\sqrt{b_{k,q}^2 + 1}\,\mathrm{Tr}(\mathbf{C}_k\mathbf{A}_{k,q}) - \frac{a_{k,q}}{\sqrt{b_{k,q}^2 + 1}}\Big) \\
& \forall k \in [1, ..., N_t], \forall q \in [1, ..., Q], \\
& C_2: \Big[\sum_{k=1}^{K} P_{cluster,k}\mathbf{v}_k\mathbf{v}^{\dagger}_k\Big]_{nn} \le P_{\text{antenna},n}, \forall n \in [1, ..., N_t].
\end{aligned} \tag{26}$$

where $\mathbf{C}_k \triangleq \mathrm{diag}(\hat{\boldsymbol{h}}_{k,q})\mathbf{Z}_k\mathrm{diag}(\hat{\boldsymbol{h}}^{\dagger}_{k,q})$, $\mathbf{Z}_k \triangleq \mathbf{W}_k - r_k\gamma_k\sum_{l\neq k}\mathbf{W}_l$, $\mathbf{W}_k \triangleq \mathbf{w}_k\mathbf{w}^{\dagger}_k = P_{\text{cluster},k}\mathbf{v}_k\mathbf{v}^{\dagger}_k$, $a_{k,q} = r_k\gamma_k N_0$, $b_{k,q} = \sqrt{2}\mathrm{erf}^{-1}(1 - 2p_{outage})$, $\mathrm{erf}(\cdot)$ denotes the Gaussian error function. The definition of $\mathbf{A}_{k,q}$ is given in the equation (22). Let $\mathbf{G}_{k,q} \triangleq \mathrm{E}\{\mathbf{E}^T \otimes \mathbf{E}_{k,q}\}$, then the elements of $\mathbf{G}_{k,q}$ is given by [35, 36]

$$[\mathbf{G}_{k,q}]_{mn} = \begin{cases} 1, & m_1 = n_1 \text{ and } m_2 = n_2, \\ \exp\{-2\sigma^2_{k,q}\}, & m_1 \neq n_1 \text{ and } m_2 \neq n_2, \\ \exp\{-\sigma^2_{k,q}\}, & \text{otherwise.} \end{cases} \tag{27}$$

where $m = (m_1 - 1)N_t + m_2$ and $n = (n_1 - 1)N_t + n_2$.

The problem P_r is complex to solve, thus, we propose a three-step low complexity algorithm for this problem. The first step of the proposed algorithm is to solve a power minimization precoding problem using semidefinite programming (SDP) [36,196] method, this power minimization problem is formulated as [36]

$$
\begin{aligned}
\mathrm{Q}: \; & \min_{\{\mathbf{W}_k\}_{k=1}^{N_t},\, t} \; t, \\
s.t.\; & \mathrm{C}_1: \|\mathbf{G}_{k,q}^{\frac{1}{2}}\mathrm{vec}(\mathbf{C}_k^H)\|_2 \; \frac{1}{b_{k,q}}\left(\sqrt{b_{k,q}^2+1}\,\mathrm{Tr}(\mathbf{C}_k\mathbf{A}_{k,q}) - \sqrt{\frac{a_{k,q}}{b_{k,q}^2+1}}\right) \\
& \forall q \in [1, ...\mathrm{Q}], \forall k \in [1, ..., N_t], \\
& \mathrm{C}_2: [\sum_{k=1}^{N_t} \mathbf{W}_k]_{nn} \le t\mathrm{P}_{\mathrm{antenna},n}, \forall n \in [1, ...N_t], \\
& \mathrm{C}_3: \mathrm{rank}(\mathbf{W}_k) = 1, \mathbf{W}_k \succeq \mathbf{0}.
\end{aligned} \tag{28}
$$

it is hard to solve the problem P with rank one solution set $\{\mathbf{W}_k\}_{k=1}^{N_t}$, we relax the rank one constraint in C_3 and use Gaussian randomization method [195, 197] given below to obtain the set of normalized precoding vectors $\{\mathbf{v}_k\}_{k=1}^{N_t}$.

$$
\begin{aligned}
& \mathbf{W}_k^{\mathrm{opt}} = \mathbf{U\Sigma U}^H, \\
& \boldsymbol{w}_k = \mathbf{U\Sigma}^{\frac{1}{2}}\boldsymbol{n}_k, \boldsymbol{v}_k = \frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|}, \forall k \in [1, ..., N_t].
\end{aligned} \tag{29}
$$

then the optimization problem can become to

$$
\begin{aligned}
\mathrm{P}_s: \; & \max_{\mathbf{r},\mathbf{P}_{\mathrm{cluster}}} \log_2(\sum_{k=1}^{N_t} r_k\gamma_k), \\
s.t.\; & \mathrm{C}_1: \|\mathbf{G}_{k,q}^{\frac{1}{2}}\mathrm{vec}(\mathbf{C}_k^H)\|_2 \le \frac{1}{b_{k,q}}\left(\sqrt{b_{k,q}^2+1}\,\mathrm{Tr}(\mathbf{C}_k\mathbf{A}_{k,q}) - \sqrt{\frac{a_{k,q}}{b_{k,q}^2+1}}\right), \\
& \forall k \in [1, ..., N_t], \forall q \in [1, ..., Q], \\
& \mathrm{C}_2: [\sum_{k=1}^{K} \mathrm{P}_{cluster,k}\mathbf{v}_k\mathbf{v}_k^\dagger]_{nn} \le \mathrm{P}_{\mathrm{antenna},n}, \forall n \in [1, ..., N_t].
\end{aligned} \tag{30}
$$

it is still complex to solve the problem P_s directly, in order to simplified the solution, we solve the power control problem ignoring the constraint C_2 in P_s

$$
\begin{aligned}
\mathrm{P}_f: \; & \max_{\mathbf{P}_{cluster}} \sum_{k=1}^{N_t} \frac{\min_{q\in[1,...,Q]} \Gamma_{k,q}}{\gamma_k}, \\
s.t.\; & \mathrm{C}_2: [\sum_{k=1}^{K} \mathrm{P}_{cluster,k}\mathbf{v}_k\mathbf{v}_k^\dagger]_{nn} \le \mathrm{P}_{\mathrm{antenna},n}, \forall n \in [1, ..., N_t].
\end{aligned} \tag{31}
$$

the objective function problem P_f is not convex, based on the definition of $\Gamma_{k,q}$, it can be transformed to problem P_m using Dinkelbach's transform [198]

$$
\begin{aligned}
\mathrm{P}_m: \; & \max_{\mathbf{P}_{\mathrm{cluster}}} \sum_{k=1}^{N_t} \min_{q\in[1,...,Q]} (\mathrm{S}_k(\mathbf{P}_{\mathrm{cluster}}) - y_m\gamma_k\mathrm{I}_k(\mathbf{P}_{\mathrm{cluster}})), \\
s.t.\; & \mathrm{C}_2: [\sum_{k=1}^{K} \mathrm{P}_{\mathrm{cluster,k}}\mathbf{v}_k\mathbf{v}_k^\dagger]_{nn} \le \mathrm{P}_{\mathrm{antenna},n}, \forall n \in [1, ..., N_t].
\end{aligned}
$$

where $\mathsf{S}_k(\mathbf{P}_{\text{cluster}}) = P_{\text{cluster,k}}\mathbf{h}^{\dagger}_{k,q}\mathbf{v}_l\mathbf{v}_l^{\dagger}\mathbf{h}_{k,q}$, $\mathsf{I}_k(\mathbf{P}_{\text{cluster}}) = \sum_{l\neq k}^{N_t} P_{\text{cluster,l}}\mathbf{h}^{\dagger}_{k,q}\mathbf{v}_l\mathbf{v}_l^{\dagger}\mathbf{h}_{k,q} + N_0$. An iterative algorithm to solve P_m has been given in [198]. $y = \frac{\mathsf{S}_k^{(m-1)}}{\gamma_k\mathsf{I}_k^{(m-1)}}$ where m stands for the iteration times. A initial feasible input $\mathbf{P}_{\text{cluster,init}}$ of P_m is required to calculate y_0. After solving P_m, the original problem P can become to

$$\mathsf{P}_u: \max_{\mathbf{r}} \log_2\Big(\sum_{k=1}^{N_t} r_k\gamma_k\Big),$$

$$s.t.\ \mathsf{C}_1: ||\mathbf{G}_{k,q}^{\frac{1}{2}}\mathrm{vec}(\mathbf{C}_k^H)||_2^2 \quad \frac{1}{b_k}\Big(\sqrt{b_{k,q}^2 + 1}\,\mathrm{Tr}(\mathbf{C}_k\mathbf{A}_{k,q}) - \sqrt{\frac{a_{k,q}}{b_{k,q}^2 + 1}}$$

$$\forall k \in [1, ..., N_t], \forall q \in [1, ..., Q], \tag{32}$$

(33)

by solving P_u, all solution of the problem P_r is obtained. The three-step low complexity algorithm is summarized in Table 10.

Table 10: Three step low complexity solution for problem P_r.

Algorithm 1
Input: $\boldsymbol{\gamma}, \mathbf{P}_{\text{antenna}}$
Variables: $\mathbf{P}_{\text{cluster}}, \boldsymbol{r}, \{\mathbf{v}_k\}_{k=1}^{N_t}$
1. Solve the problem Q to get the optimized solution $\{\mathbf{W}_k^{opt}\}_{k=1}^{N_t}$ 2. At this stage rank($\{\mathbf{W}_k^{opt}\}_{k=1}^{N_t}$) > 1 for k=1...K, do $\mathbf{W}_k^{opt} = \mathbf{U\Sigma U}^H$ Generate $\mathbf{w}_k^{opt} = \mathbf{U\Sigma}^{\frac{1}{2}}\mathbf{n}_k$ where $\mathbf{n}_k \sim \mathcal{CN}(0, \mathbf{I}_{N_t})$ Normalize $\mathbf{v}_k = \frac{\mathbf{w}_k^{opt}}{\|\mathbf{w}_k^{opt}\|}$ 3. Ignore $\mathbf{r}$, solve the objective function of P_m with the genereated $\{\mathbf{v}_k\}_{k=1}^{N_t}$, a initial feasible $\mathbf{P}_{\text{cluster,init}}$ and given $\boldsymbol{\gamma}$, $\mathbf{P}_{\text{antenna}}$ to obtain the output $\mathbf{P}_{\text{cluster}}$, to get optimal $\mathbf{P}_{\text{cluster}}$ 4. Solve the problem P_u with $\{\mathbf{v}_k\}_{k=1}^{N_t}$ and $\mathbf{P}_{\text{cluster}}$ to obtain optimized $\boldsymbol{r}$

We first examine the fairness performance of the system. Jain's fairness index can effectively quantify the total system fairness, which is expressed by $J = \frac{[\sum_{i\in\mathcal{G}}\Gamma_i]^2}{QN_t\sum_{i\in\mathcal{G}}\Gamma_i^2}$ [199], where i stands for the i-th scheduled user in the system, $J \in [0, 1]$. Higher value of J means that the SINR distribution of the system is fairer. The fairness performance of the MMSE algorithm is used to compare with the proposed algorithm, which is shown in the Figure 35. It can be observed from this figure that the fairness index of the pro- posed algorithm remains more than 0.95, which obviously outperforms that of the MMSE algorithm. Also, as the number of users per beam increases, the fairness performance of MMSE degrades remarably, however, the fairness index of proposed algorithm keeps almost unchanged.

Figure 36 depicts the outage probability of the proposed algorithm compare to that of MMSE and ZF with 10^4 channel realization, the standard deviation of the phase un- certainty is 30^{circ}. It can be observed from this figure that the outage probability of the proposed precoding scheme keeps lower than 0.01 while that of MMSE and ZF is much higher.

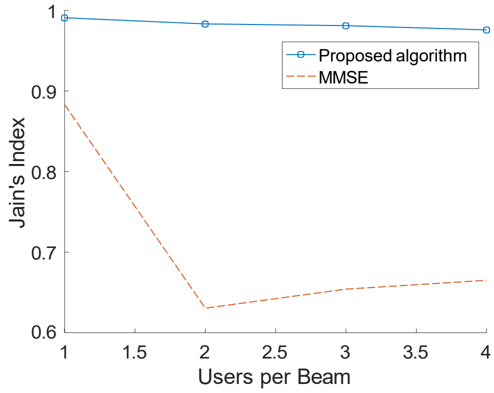


Figure 35: Jain's fairness index versus number of users per beam.

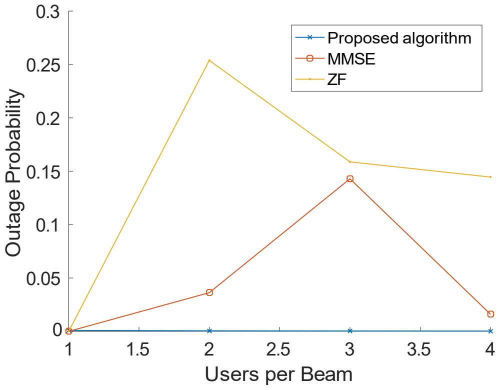


Figure 36: Outage probability as a function of users per group.

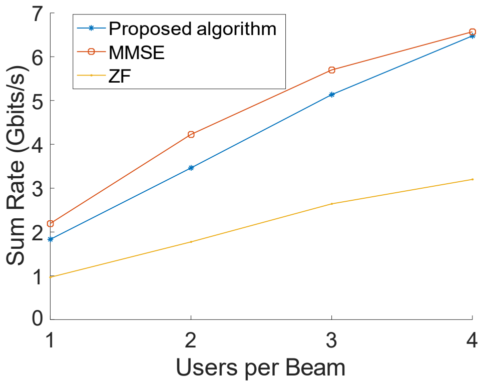


Figure 37: Sum rate versus number of users per beam.

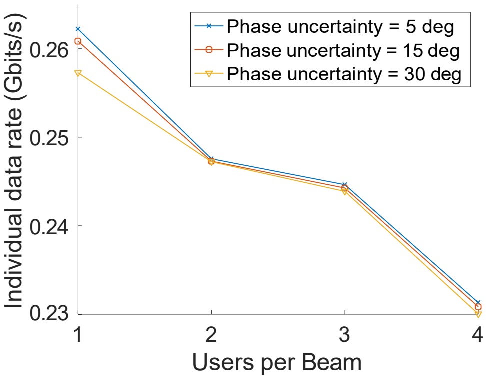


Figure 38: Average achievable data rate of individual user versus users per group with different standard deviation of phase uncertainty.

The total achievable rate of the proposed algorithm and MMSE as a function of the number of users in each user cluster is illustrated in Fig 37. A slightly sum rate degradation can be found for out proposed method due to the influence of outage constraint. This problem need more investigation in our future study.

Figure 38 demonstrates the average achievable rate of users in the satellite system as a function of the number of users in each user cluster with the standard deviation value of phase uncertainty to be 5°, 15°, and 30°. The average individual rate decreases as users per group increase as expected. We can also find that the individual rate performance degrades as the phase uncertainty increase.

10.2 Joint Precoding and User Scheduling Based on Traffic Demands

In this section, we develop a joint precoding and user scheduling approach based on traffic demands under the UFR scheme. The motivation for developing the work is as follows:

- In practical SatCom scenarios, traffic demands can remarkably vary from different users, areas, and time. Without the consideration of traffic demands, there can be undesirable cases such as insufficient or oversupply and hence lead to poor per- formance or waste of resources. As discussed in Section 2.5, precoding approaches considering traffic demands under the UFR scheme is insufficiently developed. Only a simplified scenario where each beam serves only one user was studied in existing research.

- In the existing user scheduling research, geographical positions and CSI of different users are mostly considered, and no user scheduling method based on traffic demand has been studied. As discussed in Section 2.5, balancing capacity requirements for each beam by user scheduling is significant to the performance. Hence, we develop new joint precoding and user scheduling methods that taking not only the conventional factors but also traffic demands into account.

Next, we will introduce the system model and the proposed algorithm.

10.2.1 System Model

Consider a multicast multibeam HTS with N_t transmitters generating N_t spot-beams, and the kth beam can serve a total amount of U_k users. For the multigroup multicast transmitting scheme, multiple users are served by one beam, i.e., $\sum_{k=1}^{N_t} U_k > N_t$. Considering the frame-based multicast communication, in a given time slot, the scheduler selects $card(\mathsf{G}_k) \leq U_k$ users in the kth beam and construct a codeword containing the $card(\mathsf{G}_k)$ users' data, which will be transmitted to the selected user set G_k. Users in group G_k receive the signal frame precoded by the precoding weight vector $\mathbf{w}_k \in \mathbb{C}^{N_t \times 1}$, $k \in 1, 2, \cdots, N_t$. G_k can be any subset of $\{1, \cdots, U_k\}$. We define T_k as the dictionary of all sets of G_k, i.e., $\mathsf{G}_k \subset \mathsf{T}_k$. Hence, the N_t beam is precoded by the precodingmatrix

$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N], \mathbf{W} \in \mathbb{C}^{N_t \times N_t}$, simultaneously providing service to $U_{\text{tot}} = \sum_{k=}^{N_t} U_k$ users.

Here, we assume that the perfect CSI is known at the transmitter. The channel vector between the nth user and N_t transmitters located at the kth beam can be written as $\mathbf{h}_k^{[n]} \in \mathbb{C}^{1 \times N_t}$, $n \in \{1, 2, \cdots U_k\}$, $k \in \{1, 2, \cdots, N_t\}$, which forms the kth row of the channel matrix $\mathbf{H}_n$. Let the overall channel matrix $\mathbf{H} = [\mathbf{H}^T, \mathbf{H}^T_2, \cdots, \mathbf{H}^T{}_{U_{N_t}}]^T \in \mathbb{C}^{U_{tot} \times N_t}$, and $\mathbf{s} \in \mathbb{C}^{N_t \times 1}$ be the symbol vector that contains uncorrelated and unit norm transmitted symbols, i.e., ($E[\mathbf{ss}^H] = \mathbf{I}$). After precoding, the transmitted signal vector is obtained as $\mathbf{x} = \mathbf{Ws}, \mathbf{x} \in \mathbb{C}^{N_t \times 1}$. Therefore, we can obtain the received signal vector $\mathbf{y} \in \mathbb{C}^{U_{tot} \times 1}$ as

$$\mathbf{y} = \mathbf{Hx} + \mathbf{n} = \mathbf{HWs} + \mathbf{n}, \tag{34}$$

where $\mathbf{n} \in \mathbb{C}^{U_{tot} \times 1}$ is the noise vector assuming independent and identically distributed (i.i.d.) zero-mean Additive White Gaussian Noise (AWGN).

Suppose the traffic demands for each user is known at the transmitter, the required capacity and offered capacity of the nth user at the kth beam can be written as $R_r^{(k,n)}$ and $R^{(k,n)}$, respectively. Similarly, the required and offered capacity for the kth beam are defined as $R_r^{(k)}$ and $R_o^{(k)}$.

10.2.2 Problem formulation

Assuming perfect CSI, we use $\mathbf{h}_{k,n}$ to denote the channel vector of the nth user in the kth beam. Hence, the SINR of the nth user in the kth beam is

$$\Gamma_{k,n} = \frac{|(\mathbf{h}_k^{[n]})^* \mathbf{w}_k|^2}{\sum_{j=k}^{N_t} |(\mathbf{h}_k^{[n]})^* \mathbf{w}_j|^2 + \sigma_n^2}, k, j \in 1, \cdots, N_t \tag{35}$$

where σ_n^2 is the noise power. The achievable Shannon rate is

$$R_{k,n} = B \log_2(1 + \Gamma_{k,n}), \tag{36}$$

where B is the bandwidth. With the known $R_r^{(k,n)}$, the required SINR for user n in beam k can be obtained as $\Gamma_r^{k,n} = 2^{\frac{R_r^{(k,n)}}{B}} - 1$.

Referring to the joint scheduling and precoding approach in [27], the binary variable $\eta_{k,n} \in \{0, 1\}$ is used to denote whether user n in beam j is scheduled. $\eta_{k,n} = 1$ when the corresponding user is scheduled and zero otherwise. For the kth beam, the total required and offered rate can be obtained as

$$R_r^{(k)} = \sum_{n=1}^{U_k} \eta_{k,n} R_r^{(k,n)},$$

$$R_c^{(k)} = \sum_{n=1}^{U_k} \eta_{k,n} R_{k,n} = \sum_{n=1}^{U_k} \eta_{k,n} B \log_2(1 + \Gamma_{k,n}), \tag{37b}$$

respectively.

A joint precoding and user scheduling problem can be formulated as:

$$\mathrm{P}_1 : \max_{\mathbf{W},\boldsymbol{\eta}} \sum_{k=1}^{N_t} R_c^{(k)} \tag{38a}$$

$$\text{s.t. } \mathrm{C}_1 : \eta_{k,n} \in \{0, 1\}, \forall n, \forall k, \tag{38b}$$

$$\mathrm{C}_2 : R_r^{(k)} \leq R_{\max}, \forall k, \tag{38c}$$

$$\mathrm{C}_3 : R_o^{(k)} \leq R_r^{(k)}, k = 1, \cdots, N_t, \forall k \tag{38d}$$

$$\mathrm{C}_4 : \Gamma_{k,n} \geq \eta_{k,n} \Gamma_{r,\min}^{k}, \forall k, \forall n \tag{38e}$$

$$\mathrm{C}_5 : [\mathbf{W}\mathbf{W}^H]_{kk} \leq P_k, k = 1, \cdots, N_t, \tag{38f}$$

where $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_{N_t}]$, $\boldsymbol{\eta}_k = [\eta_{k,1}, \cdots, \eta_{k,U_k}]^T$, and $\boldsymbol{\eta} = [\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_k]$. $\forall k$ refers to $k = 1, \ldots, N_t$, and $\forall n$ refers to $n \in \{1, \ldots, U_k\}$.

In (38), the optimization objective is to minimize the total unmet traffic. The con- straints C_1 to C_6 have the following insights:

- C_1: $\eta_{k,n} = 0$ can remove the corresponding item with the "suffix" k, n in the objective, C_2, and C_3, which means the user n in beam k is not scheduled.

- C_2: This constraint sets an upper bound of required rate for beam k, which avoids the case where one beam has excessive high traffic demand.

- C_3: To save resources, constraint C_3 can guarantee that the offered capacity of each beam is smaller than the required one. The constraint also makes $(R_c^{(k)} - R_r^{(k)})$ in the optimization objective larger than zero, and holds the minimization of the summation of differences.

- C_4: This constraint sets the lower bound for the achieved SINR for each user. In C_4, $\Gamma_{r,\min}^{k} = \min\{\Gamma_r^{\{k,1\}}, \Gamma_r^{\{k,2\}}, \cdots, \Gamma_r^{\{k,U_k\}}\}$. When the user is not selected, C_4 lead to $\Gamma_{k,n} \geq 0$.

- C_5: This is the conventional per beam power constraint.

10.3 Solution to the Problem P_1

Problem P is combinatorial due to the binary variable $\eta_{k,n}$, coupling with $\mathbf{W}$, which increases the complexity of solving this problem. Therefore, we propose a two-step opti- mization method to iteratively optimize $\boldsymbol{\eta}$ and $\mathbf{W}$, respectively. Firstly, an initial-value of $\mathbf{W}^{(0)}$ can be obtained through linear precoding methods such as linear minimum mean square error (LMMSE) or zero forcing (ZF) techniques.

With fixed $\mathbf{W}^{(0)}$, $\Gamma^{(0)}_{k,}$ can be obtained, and the problem P_1 can be turned into the following formation

$$
\begin{aligned}
\mathrm{P}_2:\ & \max_{\boldsymbol{\eta}} \sum_{k=1}^{N_t} \sum^{U_k} \eta_{k,n} B \log_2(1+\Gamma^{(0)}_{k,})\\
\text{s.t. } & \mathrm{C}_1: \eta_{k,n} \in \{0,1\}, \forall n,\\
& \forall k, \mathrm{C}_2: \mathbf{a}_k \boldsymbol{\eta}_k \le R_{\max},\\
& \forall k, \mathrm{C}_3: \mathbf{b}_k \boldsymbol{\eta}_k \le 0, \forall k,\\
& \mathrm{C}_4: \eta_{k,n} \le \frac{\Gamma^{(0)}_{k,n}}{\Gamma^{k}_{r,\min}}, \forall k, \forall n
\end{aligned}
\tag{39}
$$

where Constraint C_5 in (38) is temporarily left aside since it is only depends on the variable $\mathbf{W}$. In (39), $\mathbf{a}_k$ and $\mathbf{b}_k$ are defined as

$$
\begin{aligned}
&\mathbf{a}_k \text{ fj. } [R_r^{(k,1)}, R_r^{(k,2)}, \cdots, R_r^{(k,U_k)}]^T, \forall k\\
&\mathbf{b}_k \text{ fj. } [B\log_2(1+\Gamma^{(0)}_{k,1}) - R_r^{(k,1)}, B\log_2(1+\Gamma^{(0)}_{k,2}) - R_r^{(k,2)}, \cdots, B\log_2(1+\Gamma^{(0)}_{k,U_k}) - R_r^{(k,U_k)}]^T,
\end{aligned}
\tag{40}
$$

where $\Gamma_{k,n}$ can be obtained with the value of $\mathbf{W}^{(0)}$.

The combinatorial constraint C_1 can be address by relaxing it to a box constraint between 0 and 1, and implementing binary optimization methods [200].

In the second step, we will optimize $\mathbf{W}$ with the derived $\boldsymbol{\eta}^{(0)}$. Problem P_1 can be written as:

$$
\begin{aligned}
\mathrm{P}_3:\ & \max_{\mathbf{W},\boldsymbol{\eta}} \sum_{k=1}^{N_t} \sum_{n=1}^{U_k} \eta_{k,n} B \log_2(1+\Gamma_{k,n})\\
& \mathrm{C}_3: R_o^{(k)} \le R_r^{(k)}, k = 1, \cdots, N_t, \forall k\\
& \mathrm{C}_4: \Gamma_{k,n} \ge \eta_{k,n}\Gamma_{r,\min}, \forall k, \forall n\\
& \mathrm{C}_5: [\mathbf{W}\mathbf{W}^H]_{kk} \le P_k, k = 1, \cdots, N_t,
\end{aligned}
\tag{41}
$$

Constraints C_1 and C_2 are left aside as they only depend on $\boldsymbol{\eta}$. Substituting (35) into (41), we can find that the optimization objective $\max_{\mathbf{w},\boldsymbol{\eta}} \sum_{k=}^{N_t} \sum_{n=}^{U_k} \eta_{k,n} B \log_2(1 + \frac{|\mathbf{h}^H_{k,n}\mathbf{w}_k|^2}{\sum_{j/=k}^{N_t} |\mathbf{h}^H_{k,n}\mathbf{w}_j|^2 + \sigma^2_n})$ is non-convex. Therefore, following the sum rate (SR) maximization problem in [25], we introduce a slack variable $\boldsymbol{\Omega} = [\Omega_1, \Omega_2, \ldots, \Omega_{N_t}]$, with each element representing the minimum SINR for each beam, and convert the problem to the following form:

$$
\begin{aligned}
\tilde{\mathbf{P}}_3:\ &\max_{\mathbf{W},\mathbf{\Omega}} \sum_{k=1}^{N_t}\sum_{n=1}^{U_k} \eta_{k,n} B\log_2(1+\Omega_k)\\
\mathrm{C}_3:\ &\sum_{n=1}^{U_k} \eta_{k,n}[B\log_2(1+\Gamma_{k,n}) - R^{(k,n)}] \le 0,\ \forall k,\\
\mathrm{C}_4:\ &\Gamma_{k,n} \ge \eta_{k,n}\Omega_k\\
\mathrm{C}_5:\ &\Omega_k \ge \Gamma^k_{r,\min},\ \forall k,\\
\mathrm{C}_6:\ &[\mathbf{W}\mathbf{W}^H]_{ii} \le P_i,\ i=1,\cdots,N_t.
\end{aligned} \tag{42}
$$

To simplify the problem, we introduce a more rigorous constraint instead of C$_3$, as follows

$$\tilde{\mathrm{C}}_3 : \Gamma_{k,n} \le \Gamma_r^{k,n}, \forall k, \forall n, \text{if } \eta_{k,n} = 1, \tag{43}$$

which let the output SINR no greater than the required SINR for each user to avoid over supply. Substituting (35) into $\tilde{\mathrm{C}}_3$ and C$_4$, we can obtain

$$
\begin{aligned}
\tilde{\mathrm{C}}_3:\ &\mathrm{T}_1^{(k,n)}(\mathbf{W}) - \mathbf{J}^{(k,n)}(\mathbf{W}) \le 0, \forall k, \forall n, \text{if } \eta_{k,n} = 1\\
\mathrm{C}_4:\ &\mathbf{J}^{(k,n)}(\mathbf{W}) - \mathrm{T}_2^{(k,n)}(\mathbf{W}) \le 0, \forall k, \forall n, \text{if } \eta_{k,n} = 1.
\end{aligned} \tag{44}
$$

where

$$
\begin{aligned}
\mathrm{T}_1^{(k,n)}(\mathbf{W}) &\triangleq \frac{[\sum_{j=1}^{N_t} |(\mathbf{h}_k^{[n]})^* \mathbf{w}_j|^2 + \sigma_n^2]}{\Gamma_r^{(k,n)} + 1},\\
\mathrm{T}_2^{(k,n)}(\mathbf{W}) &\triangleq \frac{[\sum_{j=1}^{N_t} |(\mathbf{h}_k^{[n]})^* \mathbf{w}_j|^2 + \sigma_n^2]}{\Omega_k + 1},\\
\mathbf{J}^{(k,n)}(\mathbf{W}) &\triangleq \sum_{j\ne k}^{N_t} |(\mathbf{h}_k^{[n]})^* \mathbf{w}_j|^2 + \sigma_n^2.
\end{aligned} \tag{45}
$$

Constraints $\tilde{\mathrm{C}}_3$ and C_4 are DC. Constraints C_5 and C_6 are convex. Hence, Problem $\tilde{\mathbf{P}}_3$ is also a DC problem and can be solved by CCP. The CCP Therefore, $\mathbf{W}^{(1)}$ is derived and can be substituted into (39) to obtain $\boldsymbol{\eta}^{(1)}$.

So far, the first iteration of the two-step solution to Problem $\mathbf{P}_1$ is completed. With the value of $\mathbf{W}^{(1)}$, we can iteratively derive $\boldsymbol{\eta}^{(m)}$ and $\mathbf{W}^{(m)}$ until the stopping criterion is satisfied.

The numerical results for this joint approach is still in preparation. For the future work, we plan to further develop the joint precoding and user scheduling approach considering the time resources and adaptive power allocation.

10.4 Hybrid On-board/On-ground Signal Processing

Although the multiple gateway network can relieve the bandwidth requirements, there exist some limitations, such as the delayed CSI, and propagation variation between GWs. On-board processing [45, 46] potentially address the problems caused by the above limi- tations.

A typical hybrid space-ground precoding architecture is shown in Fig. 39 [45]. N antenna feeds form K beams, and $N > K$. With on-board beamforming, the NK fully on-ground precoding matrix $\mathbf{W}$ is transformed to $\mathbf{W}$ = $\mathbf{F}_{OB}\mathbf{F}_{OG}$, where $\mathbf{F}_{OB} \in NK$ is the on-board beamforming matrix, and $\mathbf{F}_{OG}$ KK is the on-ground precoding matrix. Clearly, the rank of on-ground precoding matrix is reduced and hence the bandwidth requirement of the feeder link.

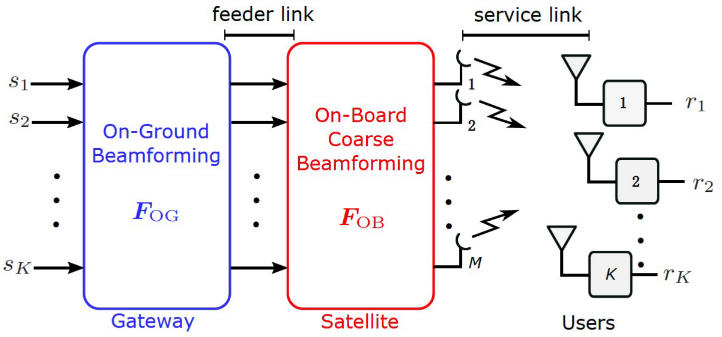


Figure 39: Block diagram of the hybrid on-board/on-ground precoding architecture. [45].

Most of the existing on on-board beam generation system [31, 46, 201, 202] is non channel adaptive (fixed), and array-fed reflectors rather than active antenna arrays are used. The fixed on-board beamforming model was also provided by the European Space Agency (ESA). Recent ten years' studies focus on how to design the fixed $\mathbf{F}_{OB}$ to mitigate the interference between feeder link and forward link, maintain fixed spot beam on-ground while compensating the satellite orbit inclination and antenna mispointing. The feeder selection has also been studied when $\mathbf{F}_{OB}$ is a sparse matrix.

Literature [203] preliminarily compared the fixed and channel adaptive on-board beam- forming using the linear MMSE beamforming approach and show the performance im- provement of on channel adaptive on-board processing. As introduced in Section **??**, with onboard APA, the design flexibility of on-board beamforming can be largely improved, and the on-board processing can become channel adaptive. Moreover, hybrid on-board/on- ground precoding share many similarities with the fully connected analog-digital hybrid precoding in cellular networks [204], and the plenty of research on analog-digital hybrid precoding offers adequate references.

Compared with the hybrid analog-digital precoding in terrestrial networks, there are peculiarities of hybrid on-board and on-ground precoding related to the nature of SatCom, as summarized as follows:

- The on-board power computational complexity requirements for hybrid on-board and on-ground precoding are more rigorous.
- The hybrid on-board and on-ground precoding deals with a much larger number of analog antennas and beams.
- The on-ground CSI is delayed, and there might be multiple gateways with limited cooperation.
- The change of on-board precoding matrix in SatCom is not as frequent as it in cellular networks.

Regarding the above characteristics of hybrid on-board and on-ground precoding and the research gap, we can see plenty of research opportunities in hybrid space-ground precoding using on-board APA.

10.5 OTFS for LEO SatCom Systems

Due to the movement of LEO satellites, precoding and user scheduling for LEO systems is quite different to GEO satellites, which is mostly considered for precoding and user scheduling techniques. This year, two interesting works [205, 206] reported a massive MIMO LEO satellite transmission scheme, under which precoding under UFR scheme was studied. These works have preliminarily established the feasibility and appealing benefits of implementing UFR precoding and user scheduling in LEO satellite systems. In these works, the authors established the massive MIMO channel model for LEO and proposed linear precoding and user scheduling methods under this massive MIMO model. They further developed linear precoders with the use of uniform planar arrays (UPAs) onboard the satellite. Simulation results in [205] demonstrated that with the use of FFR and precoding techniques, significant sum-rate gains can be observed over the conventional 4CFR approaches.

Besides, various Doppler shift estimation algorithms for LEO have been proposed in [3, 4, 207], which also lay a foundation of channel estimation, precoding and user scheduling in LEO. Due to the high speed of LEO satellites, the delay-Doppler domain based on principles of new Orthogonal Time Frequency and Space (OTFS) scheme is very promising, in order to combat high Doppler frequency shift and provide Doppler-resilient performance in a very dynamic changing environment. Next, we will review the OFTS technology and show some simulation results.

10.5.1 Fundamentals of OTFS

OTFS modulates information in the delay-Doppler (DD) domain rather than in the con- ventional time-frequency (TF) domain of classic OFDM modulation, which results in delay- and Doppler-resilience, whilst enjoying *joint time-frequency diversity* (termed as *full diversity* in [208]), which is the key for supporting reliable communications. Addi- tionally, OTFS modulation has the potential of transforming a time-variant channel into a 2D quasi-time-invariant channel in the DD domain, where its attractive properties can be exploited.

I. From Time-Invariant to Time-Variant Channels

Wireless channels can be modeled by a linear time-invariant (LTI) system, provided that the channel impulse response (CIR) is time-invariant or has a long coherence time. In the presence of multiple scatters, the dispersive LTI channel's output is a temporally smeared-out version of the transmit signal, but again, the CIR is *time-invariant*. In this case, a one-dimensional (1D) CIR in the delay domain $h(\tau)$ is sufficient for characterizing the time-dispersive channel. The Fourier transform (FT) of this CIR is a frequency- selective channel transfer function (CTF). As the delay spread of the CIR is increased, the selectivity becomes more severe, since the separation of the frequency-domain (FD) fades is increasingly proportional to the CIR-length.

However, the assumption of having LTI CIRs may no longer hold in the face of in- creased user mobility and carrier frequency. Therefore, the linear *time-variant* (LTV) channel model [209] has attracted considerable research attention in high-mobility sce- narios. LTV channels give rise to frequency shifts due to the Doppler effect, yielding a spectrally smeared version of the transmitted signal, i.e., they are frequency-dispersive. Frequency-dispersive channels are *time-selective* and the separation of the channel's time- domain (TD) fades is increasingly proportional to the Doppler spread. In practice, the

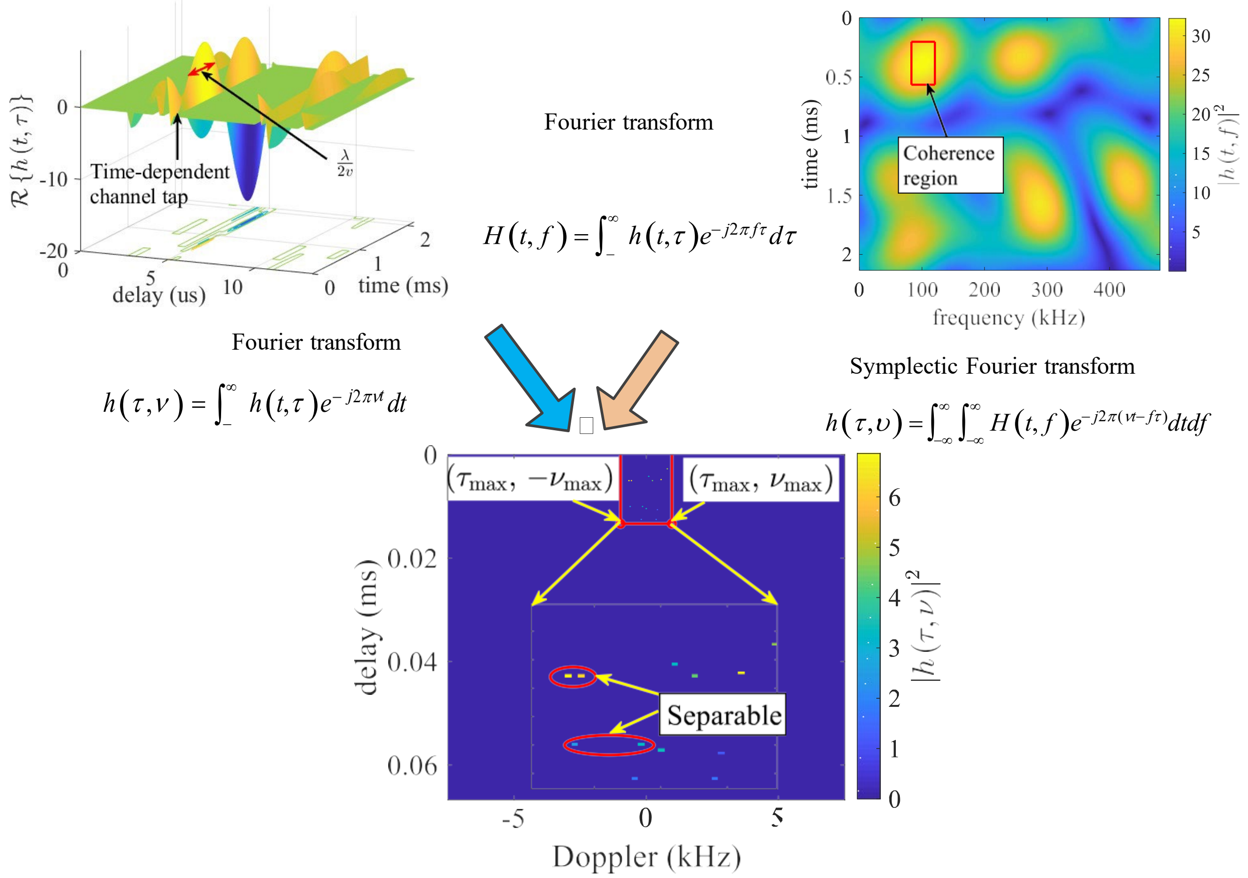


Figure 40: LTV channels in the time-delay, TF, and DD domains.

LTV channels of high-mobility scenarios are often *doubly-dispersive* due to the joint pres- ence of dispersive multipath propagation and the Doppler effects. The transmitted signals suffer from dispersion both in the TD and FD. In such scenarios, each tap of the CIR function is time-dependent, fluctuating according to the rate of $\frac{\lambda}{2}$ between consecutive TD fades, as shown in Fig. 40, where λ denotes the wavelength and υ is the relative speed between the transmitter and receiver. Hence, this results in a 2D CIR function $h(t,\tau)$ in the time-delay domain. In contrast to the traditional way of treating TD and FD disper- sion as undesired channel impairments, we can beneficially exploit the additional degrees of freedom (DoF) of these channels for achieving reliable diversity-aided communications over high-mobility channels.

II. LTV Channels in TF and DD Domains

Apart from the time-delay domain channel $h(t,\tau)$, the LTV channels can be equiva- lently described in either the TF or DD domain, as shown in Fig. 40. To emphasize the TF selectivity, the TF domain channel, $H(t,f)$, can be obtained by the FT of $h(t,\tau)$ with respect to (w.r.t.) the delay τ. Note that $H(t,f)$ can be interpreted as the com- plex CTF coefficient at time instant t and frequency f. Due to the limited coherence time and coherence bandwidth (coherence region in Fig. 40) of LTV channels, channel state information (CSI) acquisition in the TF domain would be challenging and would impose a significant signaling overhead. For instance, for an OFDM system having a carrier frequency of $f_c = 3.5$ GHz and a subcarrier spacing of $\Delta f = 15$ kHz supporting a relative velocity of $\upsilon = 300$ km/h, the maximum Doppler shift is $\upsilon_{\max} = 972.22$ Hz and

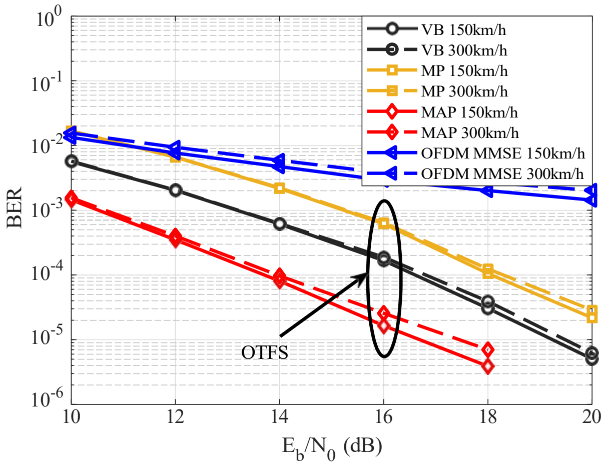


Figure 41: Performance comparison of OTFS with different detectors, moving speeds, and waveforms. The number of paths in the DD domain is 4, $N = 8$, $M = 16$, $\upsilon_{\max} = \frac{3}{NT}$ for the velocity of 150 km/h, $\upsilon_{\max} = \frac{6}{NT}$ for the velocity of 300 km/h, and $\tau_{\max} = \frac{3}{M\Delta f}$.

the OFDM symbol duration including a 20% cyclic prefix (CP) is 80 μs. Assuming that the channel's coherence time is $\frac{1}{4}_{\max}$ = 257.14 μs, the channel's coherence interval can only accommodate at most 3 OFDM symbols.

Applying the FT to $h(t, \tau)$ w.r.t. t yields the DD domain channel (spreading function), $h(\tau, \upsilon)$. The DD domain channel $h(\tau, \upsilon)$ characterizes the intensity of scatters having a propagation delay of τ and Doppler frequency shift of υ, which directly captures the underlying physics of radio propagation in high-mobility environments. More importantly, the LTV channel in the DD domain exhibits the beneficial features of separability, stability, compactness, and possibly sparsity, as illustrated in Fig. 40, which can be exploited to facilitate efficient channel estimation and data detection.

10.5.2 Efficient DD Domain Data Detection in OFTS Systems

As a fledgling waveform, OTFS modulation unveils new opportunities but also has its own challenges. In this section, we will discuss the DD Domain Data Detection problem. As shown in [210], the output signal in the DD domain can be regarded as a 2D circular convolution of the input data symbols and the effective aggregate channel, which results in a rather specific interference pattern, where a pair of symbols far from each other in the DD domain may interfere with each other. Mitigating this peculiar interference requires a bespoke receiver. Adopting the optimal maximum a posteriori (MAP) detector would indeed perfectly mitigate the interference between symbols, but at an excessive complexity, precluding its deployment in practical systems. Hence, most OTFS detectors focused on the complexity reduction, based on the classic message passing algorithm (MPA) and its variants. The main problem of MPA-based detection is its poor convergence behavior in the face of short cycles, which may lead to performance degradation.

A potent solution is to adopt the variational framework of [211], which can adaptively construct the distributions of OTFS symbols according to their interference patterns. By appropriately constructing the distributions of OTFS symbols for variational purposes, we can design rapidly converging OTFS detection. An initial result adopting the variational Bayes (VB) OTFS detector achieves a modest performance gain over the MPA owing to its better convergence, as depicted in Fig. 41. The performance of MAP detection is also provided as the baseline, which has the best performance, albeit at the cost of an excessive complexity. For all these detectors, the OTFS performance remains similar upon increasing the velocity from 150 km/h to 300 km/h. On the other hand, the detection performance of the minimum mean squared error (MMSE) OFDM detector remains poor due to the excessive inter-carrier interference (ICI).

10.5.3 Potential Opportunities of Applying OTFS to SatCom Systems

As shown in Table 1, the next-generation SatCom systems will operate in higher frequency band, such as Ku and Ka. The Doppler effect becomes more severe upon increasing the carrier frequency even at a low/medium velocity, not to mention the LEO satellites with extremely high velocities, e.g., the mean velocity is 28,080 km/h. Although increasing the subcarrier spacing to mitigate the resultant ICI is feasible, the TD symbol duration will be shorter and inserting a CP for guarding against ISI will introduce a significant overhead. The excessive phase noise associated with high-frequency oscillators also results in a time-varying composite channel. OTFS provides strong immunity to the oscillator phase noise, which is crucial for mmWave communications.

Non-terrestrial networks (NTN) are capable of supporting the terrestrial 5G networks in the provision of global coverage and mobility, as well as ubiquitous connectivity and enhanced network reliability. The high Doppler spread experienced by the NTN, partic- ularly for the NGSO systems imposes new challenges on its air interface design. OTFS modulation has rich potential in the NTN owing to its prominent capability of handling the Doppler effect. Additionally, satellites have limited on-board power supply and com- puting capability, and hence the low PAPR and low complexity of OTFS is of pivotal importance. Moreover, the corresponding NTN communication links spanning to the ground terminals usually exhibit spatial channel sparsity in the DD domain, which allows OTFS to strike an attractive performance vs. complexity trade-off.

10.6 Conclusions

In this section, we mainly present four parts of our research for the next-generation HTS systems, i.e., the robust outage constrained proportional fairness algorithm, joint precoding and user scheduling based on traffic demands, OTFS for LEO SatCom Systems, and hybrid On-board/on-ground signal processing techniques. For each part, the key findings are as follows:

- Towards the phase uncertainties of CSI, we proposed a robust precoding algorithm that guarantees the fair resource allocation among users. The preliminary simulation results verify its robustness to phase uncertainties, as well as the total system fairness. The proposed approach successfully solves a practical precoding problem with low complexity. In the next phase, user scheduling method will be jointly studied based on the current precoding algorithm.
- Balancing capacity requirements for each beam by user scheduling is significant to the performance of precoding. Therefore, for the first time, we propose joint precoding and user scheduling approaches based on traffic demands under the multigroup multicast UFR scheme. The capacity requirements for each user and beam is re- garded as a new criterion for user scheduling methods. We also formulate new joint

precoding and user scheduling matching traffic demands problem and have success- fully found a way of solving the original non-convex and combinatorial problem.

- Existing studies on hybrid on-board/on-ground precoding are insufficient on channel adaptive on-board processing, which has better performance than the fixed on-board beamforming. Moreover, there is much potential in implementing hybrid on-board/on-ground precoding with onboard APAs, which shares many similarities with hybrid analog/digital precoding in terrestrial networks, and can significantly improve the on-board processing flexibility. However, to our best knowledge, there is no research on such novel hybrid systems with on-board APA. Taking adequate research on hybrid analog/digital precoding in cellular networks as excellent potential references, and be mindful of the peculiarities of SatCom systems, there are numerous research opportunities in this area.

- OFTS, a new modulation technology in its infancy, owes many appealing features in combating high Doppler frequency shift. This is particularly beneficial for the next-generation NGSO SatCom systems with high-speed satellite movements. In this section, we introduce fundamentals of OTFS, and explicit its advantages over the existing modulation schemes in terms of Doppler-resilience with theoretical analysis and simulation results. It is shown that OTFS has great potential for LEO satellites with frequent and high-speed movements, and its low PAPR and low complexity is also desirable for SatCom systems.

11 Proof of Concept Implementation of Machine Learning Algorithms to Handle Satellite Channel Impairments

In this section, we present a general high power amplifier (HPA) model with memory effect, and propose a novel neural network (NN)-based digital pre-distortion (DPD) to combat the nonlinearity in HPA. The nonlinearity is an inherent characteristic of amplifiers and is one of the most challenging problems in satellite communications. Typically, the nonlinearity will result in adjacent channel leakage, degraded error performance, and force the transmitter to reduce its transmission power into a more linear but less power-efficient region [105]. Moreover, HPAs also suffer from memory effects, which are caused by time variations in the amplifier's circuit characteristics (e.g., the charging and discharging of capacitors and inductors), which further exacerbates the situation when dealing with the nonlinearity. Digital pre-distortion (DPD) is a widely used baseband signal processing technique to improve the linearity of the radio transmitter at an HPA. Before the signal is distorted by the HPA, it will be pre-processed by DPD in order to counteract the nonlinearity of the HPA.

11.1 High Power Amplifier (HPA) Model with Nonlineary and Memory Effect

In this subsection, we present an HPA model considering both the nonlinearity and memory effects. We model the HPA as a discrete-time system. The model processes the baseband signals and ignores the noise and ambient factors. As shown in Fig 42, the HPA model consists of a Saleh amplifier [212] which distorts the input signals, together with a finite impulse response (FIR) filter which can be viewed as a linear time-invariant (LTI) system to characterise the memory effect. In this model, we also define an HPA backoff level to indicate how close the HPA is driven to saturation.

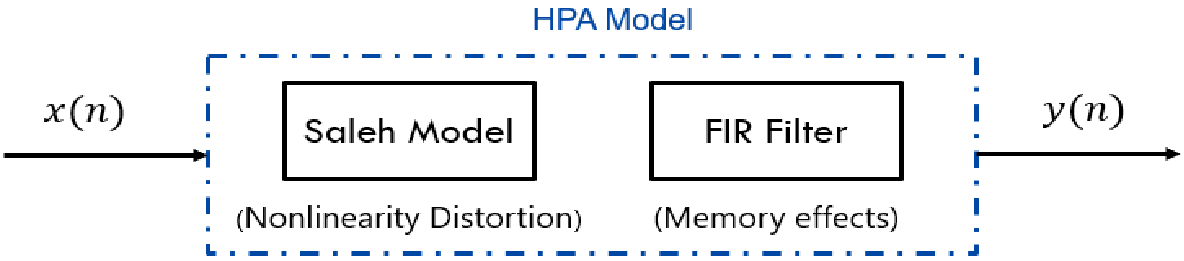


Figure 42: The HPA model with nonlinearity and memory effect

Basically, the signal can be distorted in both amplitude and phase shift. Mathematically, the Saleh model distorts signal as

$$F_a(\mu) = \frac{\alpha_{am}\mu}{1+\beta_{am}\mu^2}, F_{ph}(\mu) = \frac{\alpha_{ph}\mu}{1+\beta_{ph}\mu^2}, \quad (46)$$

where μ is the magnitude of input signal, $F_{am}(\mu)$ and $F_{ph}(\mu)$ are the functions of amplitude and phase shift behaviors for the Saleh model, respectively. α_{am}, β_{am}, α_{ph}, and β_{ph} are all the model parameters.

From a system perspective, the HPA has a system function $H(z)$, and the DPD system function is $G(z)$. Our objective is to find an appropriate $G(z)$ so that $G(z)H(z)$ = 1. That is, for any input signals, we can get similar signals at the output.

A sum of polynomial functions are usually used to characterize the HPA performance [213], which can be shown as

$$y_n = \sum_{k=1}^{K} \sum_{q=0}^{Q} a_{kq} x_{n-q} |x_{n-q}|^{k-1}, \qquad (47)$$

where n is the time index, x_n and y_n are input and output signal of the HPA at time n, respectively, a_{kq} is the HPA polynomial coefficient where q is HPA memory depth, and k is the degree of HPA nonlinearity. This model performs well if the HPA is not complex or does not consider the memory effect. However, when the nonlinearity of HPA is severe, the degree of the polynomial needs to be increased in order to maintain accurate modeling.

11.2 The Proposed Neural Network Based Digital Pre-Distortion Scheme

In this subsection, we propose a novel low-complexity NN-based DPD scheme to combat the nonlinearity of the HPA when considering the memory effect. The diagram of the proposed scheme is shown in Fig. 43. We use the strong mapping capability of NN

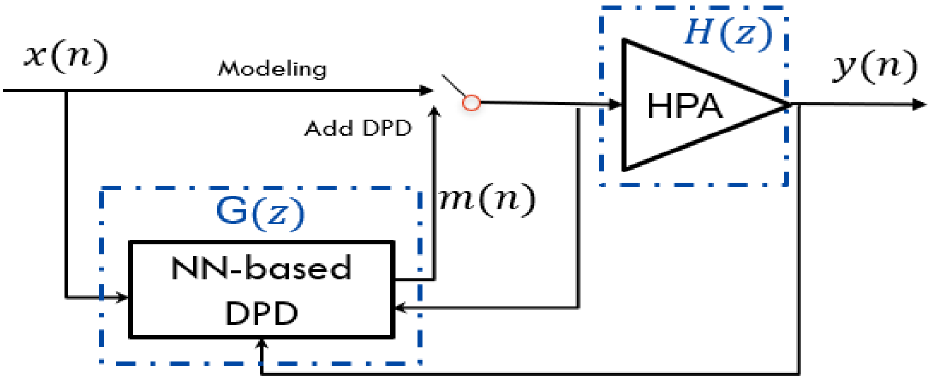


Figure 43: Block diagram of a HPA with NN based DPD models

to capture the underlying complex relationship between the input (original signal) and output (distorted signal) of the HPA model.

We first give an example of a single hidden layer based NN in Fig. 44. A single hidden layer NN is a function f: R^D R^L, where D is the size of input vector and L is the size of the output vector. Specifically, we have $f(\vec{x}) = S_2(b^2 + W^2(S_1(b^1 + W^1\ x)))$, where x, $f(x)$ are the input and output, respectively, b^1, b^2 are bias vector, W^1, W^2 are weight matrices, and S_1, S_2 are activation functions such as tanh function $\tanh(a) = (e^a\ e^{-a})/(e^a + e^{-a})$ and sigmoid function $sigmoid(a) = 1/(1 + e^{-a})$. To train a neural network, we need to learn the parameters of the model $\theta = W^1, b^1, W^2, b^2$.

Next, we build up a fully connected neural network as shown in Fig. 45 to represent the pre-distortion in DPD. There are two input and output neurons for the real part

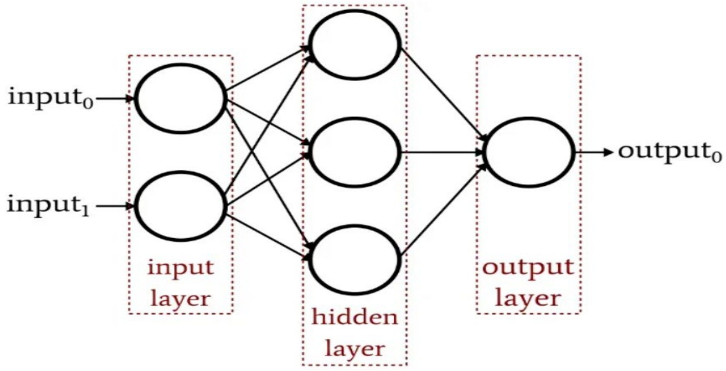


Figure 44: A example of single hidden layer neural network.

Figure 45: General structure of the DPD neural networks

and the imaginary part of the signal and three hidden layers with 6, 8, and 6 neurons, respectively. The hidden layer function is defined as

$$h(x^j) = S_j(b^j + W^j \times x^j),\ j = 1, 2, 3 \tag{48}$$

where x^j, b^j, W^j, and S_j are input data, bias vector, weight matrices, and activation function for j-th hidden layer, respectively. We use the sigmoid function as the activation function. If x denotes the input data, then the output data can be shown as

$$f(x) = S_4(b^4 + W^4 \times S_3(b^3 + W^3 \times S_2(b^2 + W^2 \times S_1(b^1 + W^1x))))), \tag{49}$$

where b^4, W^4, and S_4 are the bias vector, weight matrices, and activation functions for the output layer, respectively. We aim to minimize the following loss function

$$\min L = \frac{1}{M}\sum_{i=1}^{M} |f(a_i) - b_i|^2, \tag{50}$$

where M represents the number of signal samples, a_i, b_i represent the i-th input sample and labeled output sample, respectively, and $f(a_i)$ is the output data shown in (49).

11.3 Preliminary Results

The dataset are collected from the HPA model in Subsection 11.1 on the Matlab Simulink platform. We create a random binary data stream and map the data stream to a 16- QAM constellation in signal carrier system, and then it used as the input data of the HPA. We set the HPA backoff level as 7 dB, and set $[\alpha_{am}, \beta_{am}] = [2.1587, 1.1517]$, $[\alpha_{ph}, \beta_{ph}] = [4.0330, 9.1040]$ in (46). The input data and the output labeled data of our NN- based DPD are $y(n)$ and $x(n)$ of the HPA model in Fig. 42, respectively. We collect 40000 samples and divide them into two parts: 70% training set and 30% testing set. For the NN in Fig. 45, the initial bias vectors and weight matrices are randomly generated. The learning rate $\tau = 0.001$ and the total number of epochs $\eta = 300$. Once the NN is well trained, if we input the original data $x(n)$, we will get the pre-distorted data $m(n)$ before entering the HPA.

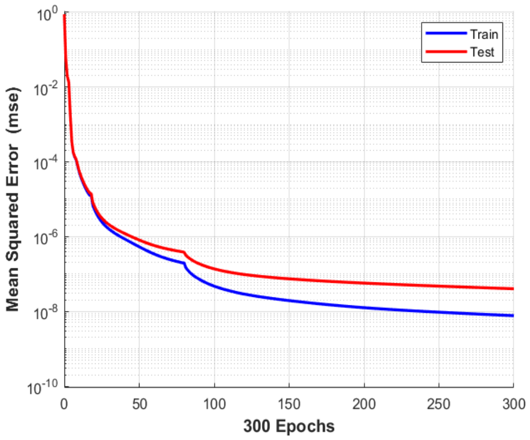
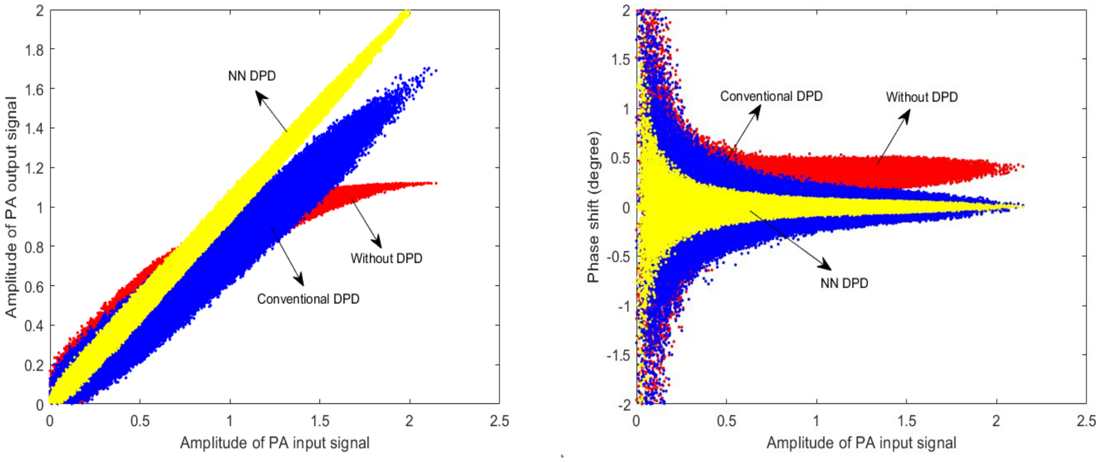


Figure 46: Mean square error (MSE) performance of the NN.



(a) Amplitude performance

(b) Phase shift performance

Figure 47: Comparison between different output scenarios: (a) Amplitude distortion. (b) Phase shift.

Fig. 46 shows the mean squared error (MSE) performance of the NN in terms of the number of epochs. It can be shown that the convergence speed is very fast, i.e., we have MSE = 7.848 10^{-9} when the number of epochs is 300.

In Fig. 47, we compare the performance of our proposed NN-based DPD (yellow dots) with the conventional DPD method (blue dots) [212] and the original method without the pre-distortion process (red dots) in terms of amplitude (Fig. 47(a)) and phase shifts (Fig. 47(b)) between output and input signal. In Fig. 47(a), it is clearly shown that the NN-based DPD significantly outperforms the conventional DPD in linearizing the amplitude of the output. Fig. 47(b) demonstrates that the NN-based DPD scheme also outperforms the conventional DPD in decreasing the phase shift between the output and input data.

11.4 Conclusion and Future Directions

The proposed NN-based DPD scheme can largely mitigate the nonlinearity of HPA. It can characterize the HPA’s nonlinear behavior and the memory effect. By pre-processing the signal before HPA, it can linearize the amplitude gain and reduce the phase shift of output of HPA. NN-based DPD has a better performance compared to conventional DPD. In the future, we want to reduce the number of nodes and layers of NN, and it is also possible to include memory cells such as recurrent neural networks (RNNs) in the NN to account for memory effects. Besides, we plan to adopt the environmental factors to build a more general HPA model with time-varying parameters. From a system perspective, we plan to validate the effect of NN-based DPD on other parts of the communication system such as orthogonal frequency division multiplexing (OFDM), multiple-input and multiple-output (MIMO).

12 Proof of Concept Implementation of FPGA-Based Fault Tolerant SDR for the Satellite Transponder

This part of the project aims to use FPGA and artificial intelligence to achieve an SDR implementation for 5G protocol. The first step is to simulate 5G baseband in MATLAB and then to design an SDR system with CPU to implement the baseband. The research focus will be the hardware and software co-design architecture. By using FPGAs, the compute intensive tasks including digital signal algorithms, data connections, and fault correction algorithms can be done in parallel and pipelined behaviour. The resources and clock cycles can be used more efficiently to improve the overall performance while limiting power consumption.

12.1 Model-based Design

To achieve the FPGA-based SDR system, a reliable method of converting software algorithms to hardware has to be used. MathWorks provides a model-based design to accelerate prototypes and 5G field trials [214]. Firstly, the developers can design software algorithms and test bench in MATLAB. Then fixed-point algorithm blocks are used to develop a hardware-friendly model in Simulink. The model can be divided into several sub-modules to target the FPGA. HDL and C code can be generated automatically by the software and then prototype on SDRs hardware. Finally, the full radio platform design will be integrated and tested.

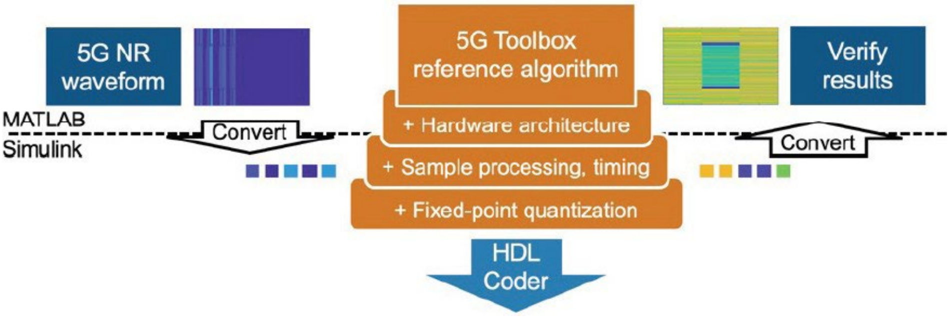


Figure 48: Top-down refinement workflow from wireless algorithm to FPGA deployment.

Fig. 48 gives a workflow overview to convert from a software algorithm to an FPGA [215]. In 5G SDR system, reference algorithms are the baseband components including digital signal processing algorithms. They will be developed and tested in MATLAB and work as the “golden reference” for the following steps. To convert the algorithms for hardware implementation, the different functionalities in the MATLAB code need to be partitioned to sub-modules, and inputs and outputs should be defined for each sub- module. In this step, software and hardware based functions have to be distinguished. Normally, the software is used to control the hardware acceleration in the hardware- software co-design wireless applications. For hardware-based algorithms, Simulink can be used to design the sample-based processing and timing to ensure the proper operation with streaming sample data. Parallel and pipelined design structure are distributed in the Simulink model. However, a trade-off between resource usage and algorithm performance should be considered as well. Fixed-point implementation is another important part of digital hardware designs. A better use of fixed-point operations can significantly improve efficiency while maintaining accuracy. Many calculation methods based on fixed-point numbers can be used to solve complex operations. For example, coordinate rotation digital computer(CORDIC) is commonly used in FPGA development to calculate trigonometric functions, square roots, and exponential as it only requires shift-and-add algorithms. After the model design and hardware-related optimization, HDL Coder will generate VHDL or Verilog HDL to target the FPGA devices. In addition, the generated HDL code can be further improved by the developers.

12.2 Example analysis and implementation

A built-in 5G new radio example is being analyzed and tested to perform the model-based design. The 5G NR Cell Search design is used to detect and demodulate synchroniza- tion signal blocks (SSBs). Detected Primary and secondary synchronization signals (PSS and SSS) will be used to obtain the initial system information from a gNodeB(gNB) Sig- nal [216]. Fig. 49 shows the hardware and software partition of the algorithm, where the Search Controller and SSB Detector will be deployed to software and hardware separately. The Digital Down-Converter (DDC) block performs frequency translation to correct fre- quency offsets in the received signal waveform. The PSS search block searches for SSBs and returns a list of PSSs to the software so that it can filter out the strongest cell ID. The Orthogonal Frequency Division Multiplexing (OFDM) demodulation block demodulates an SSB resource grid and also the strongest PSSs. Then, the SSS search block searches for SSS and outputs the overall cell ID.

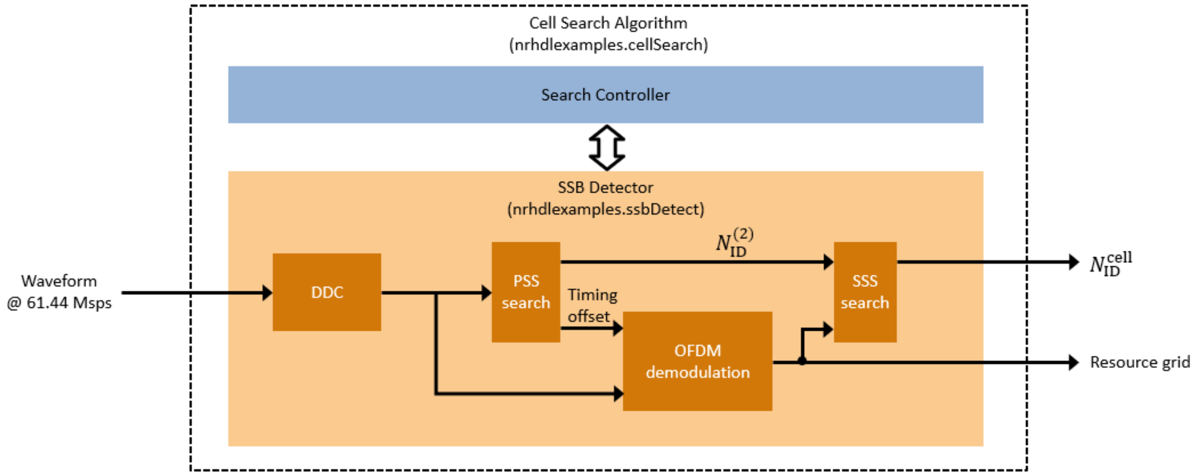


Figure 49: The 5G NR HDL Cell Search algorithm.

The example provides the reference algorithm including waveform generation, channel propagation, and the receiver algorithms in one script as the first step of the workflow. To implement the algorithm on hardware, the receiver functionalities and inputs and outputs have to be partitioned into individual MATLAB functions as sub-modules. The Simulink model is generated from the hardware algorithm reference to manage the memory allocation, design the parallel and pipelined processing structure, and also adjust the fixed- point bit widths and operations. Fig. 50 gives an overview of the SSB Detection model structure.

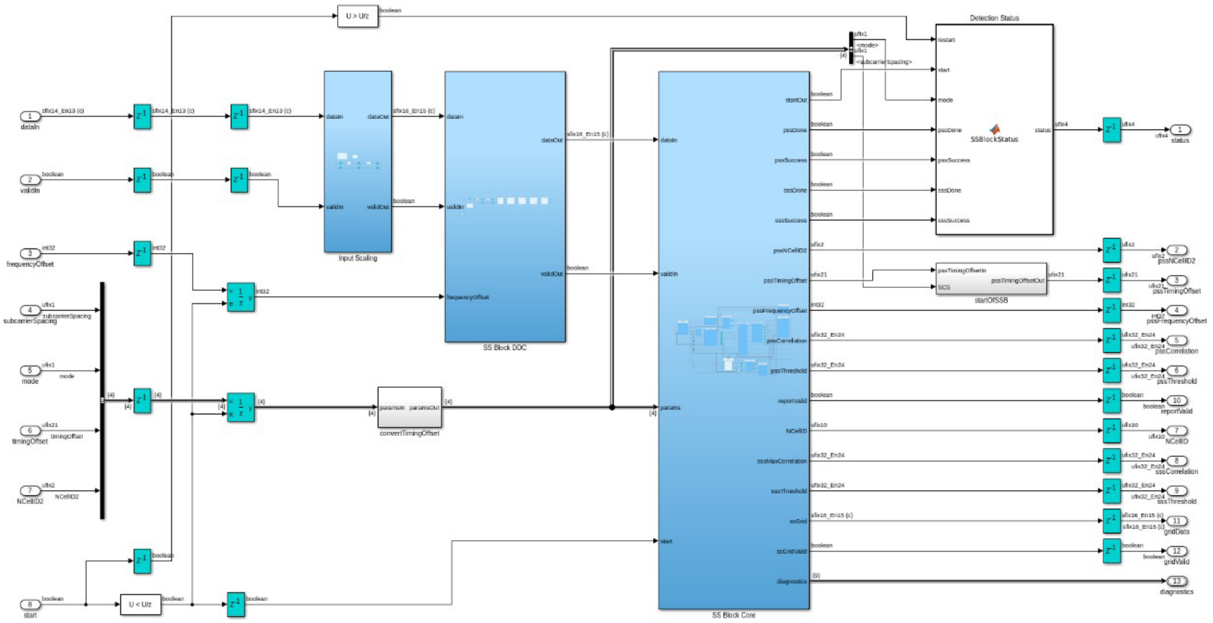
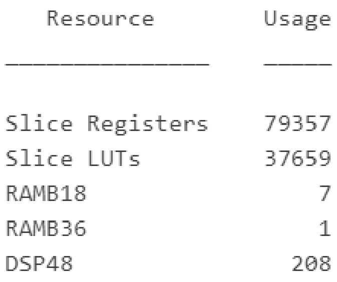


Figure 50: SSB Detection Model Reference Structure.

After that, HDL Coder will be used to generate VHDL or Verilog HDL to target the selected hardware device. A further hardware-related optimization can be applied to the generated HDL code as well. The final Cell Search model can be successfully synthesized and implemented for a Xilinx Zynq-7000 FPGA board with 230 MHz clock frequency and Fig. 51 gives the overall resource usage of the system.



```
    Resource        Usage
_______________     _____

Slice Registers     79357
Slice LUTs          37659
RAMB18                  7
RAMB36                  1
DSP48                 208
```

Figure 51: Hardware resource utilization of the Cell Search algorithm.

12.3 Conclusion and Future Directions

In summary, it is a feasible solution to convert software algorithms to a hardware and software co-design system. The next steps include converting 5G-related SDR functions to FPGA-based algorithms, optimizing the algorithms by maximizing the advantages of FPGA, and verifying and debugging the system. Fault tolerance will be considered for each sub-module separately, and then for the overall system performance. Radia- tion interference will be the main problem to be solved for fault-tolerance to ensure the proper use of the proposed system in satellite transceivers. There are many available solutions including using radiation-tolerant devices, adding error correction codes, us- ing over-voltage protection circuit and current-sensing monitoring network, and also the triple-mode-redundancy.

13 Proposed Artificial-Intelligence-Enabled SDN Architecture for 5G NTNs

We propose an artificial-intelligence (AI) enabled SDN architecture for 5G NTNs as shown in Fig. 52. In this architecture, we segment the 5G NTN into three parts: 1) a 5G NTN core network located in the data centres around the globe, 2) a 5G NTN radio access network including the ground stations, satellites and user equipment, and 3) a 5G NTN satellite networks in the space. To enable programmability of the whole network, we need to implement agents in network devices, where each agent controls the network device and communicates with the AI algorithms running in the data centres. With such an architecture, we can optimise the end-to-end performance across the whole networks by designing a joint algorithm controlling three different parts of 5G NTNs.

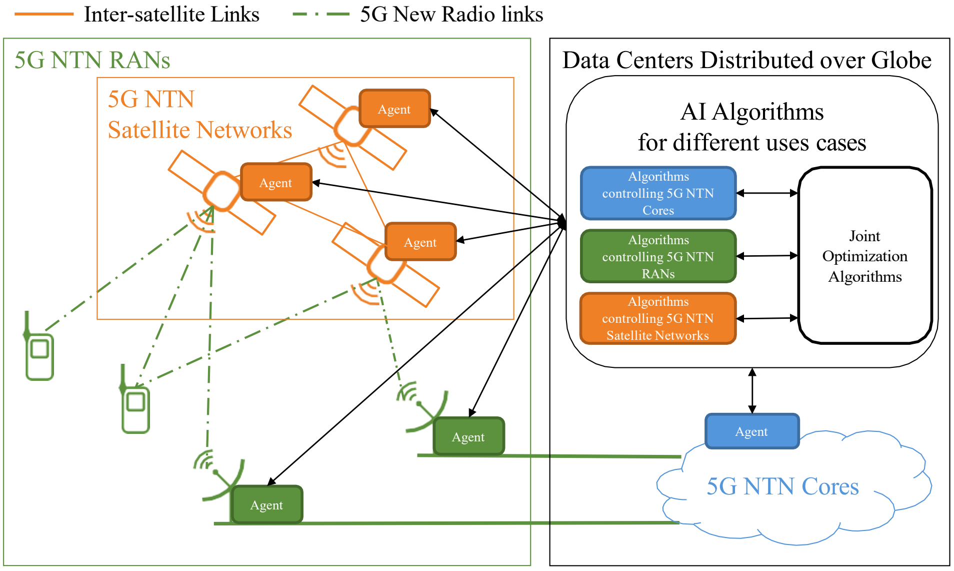

Inter-satellite Links
5G New Radio links
5G NTN RANs
5G NTN
Satellite Networks
Agent
Agent
Agent
Agent
Agent
Data Centers Distributed over Globe
AI Algorithms
for different uses cases
Algorithms controlling 5G NTN Cores
Algorithms controlling 5G NTN RANs
Algorithms controlling 5G NTN Satellite Networks
Joint Optimization Algorithms
Agent
5G NTN Cores

Figure 52: AI-enabled SDN architecture for 5G NTN.

In order to bring the above architecture into reality, we decompose the research prob- lem into three parts as shown in Fig. 53. First of all, it is important to note that the 5G NTN is not yet standardised, though 3GPP standardisation activities are underway. Technical problems will occur when migrating 5G from terrestrial networks to NTNs. Some of these technical problems have been identified by 3GPP in [176] along with their possible solutions. Efforts could be made to evaluate these solutions or further propose new ones. This can be done by mathematically formulating the relevant optimization problems in these issues and then developing the optimal solutions for them.

After we develop a clear view on 5G NTN standards, the next step is to separate the logic of decision making, referred to as the control plane, from the logic of data processing, referred to as the data plane in 5G NTNs. The control plane will be centrally located in the data centre, in which the AI algorithms can manipulating the networks. The data plane remains in the networks devices of 5G NTNs. The agent in the network device bridges the communication between the data plane and the control plane. The communication interfaces between the agent and the data plane, and the interfaces between the agent and the control plane are referred to as southbound and northbound, respectively.

On the southbound link, the states of the network device come inwards to the agent and the agent outputs controlling decisions to the network device. For the northbound link, there are two types of configurations for different SDN applications. The first type of the northbound link is where the agent makes a query to the control plane in order to get a controlling decision for the given state, for example, Openflow protocol [85] and FlexRAN interface [169]. Note that this northbound will cause a delay in the controlling decision. As a result, it is only suitable for a non-real time application. For real time applications, we can implement the second type of the northbound link where the control plane distributes the logic of decision making to the agent [184]. Then, the distributed logic in the agent takes the state of the network device as the input and returns the controlling decision as the output. To continuously improve the logic, the agent streams the state-decision pairs to the control plane, and the AI algorithm evaluates these pairs and optimises the logic automatically. An example of this type of the northbound link can be found in [184] where a neural network (NN) is trained in a central server by an AI algorithm and is distributed to 5G base station as the radio resource scheduling logic.

Once the SDN architecture is developed, we will have the ability to develop AI algo- rithms in the data centres in order to control and optimise 5G NTNs. To achieve this, we need first formulate optimal control problems for different SDN applications. Then, we can develop AI algorithms to solve these problems. Finally, the developed algorithms can be deployed in the data centres.

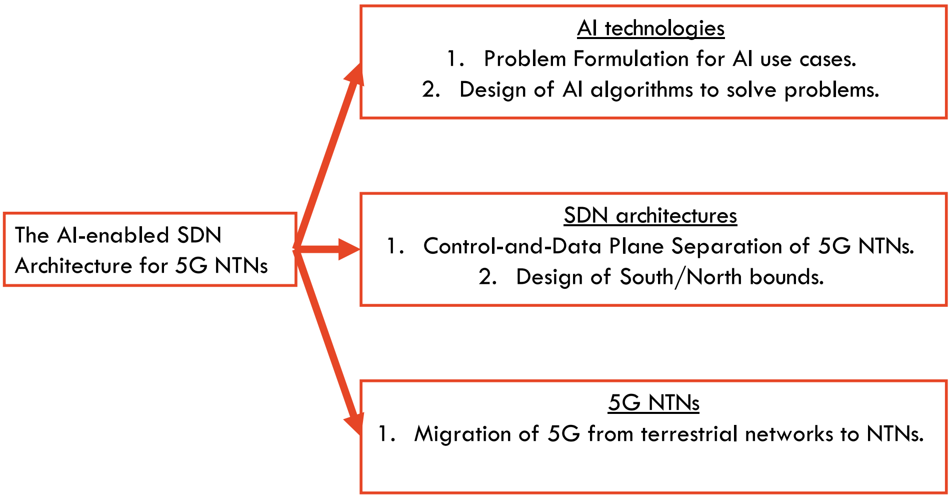

The AI-enabled SDN Architecture for 5G NTNs
AI technologies
1. Problem Formulation for AI use cases.
2. Design of AI algorithms to solve problems.
SDN architectures
1. Control-and-Data Plane Separation of 5G NTNs.
2. Design of South/North bounds.
5G NTNs
1. Migration of 5G from terrestrial networks to NTNs.

Figure 53: Decomposition of the research problem on AI-enabled SDN architecture for 5G NTNs.

13.1 Preliminary Evaluation

13.1.1 System model

We consider a forward link of 5G NTN in the transparent architecture, as shown in Fig. 54, where a GEO stationary satellite amplifies and does a carrier frequency shift on the signal transmitted by a ground station and sends the processed signal on the downlink to users. The elevation angles of the users and the ground station are 50° and 90°, respectively. Shadowing in the downlink is modelled by a log-normal distribution with a zero mean and a standard deviation of 2.7 dB. We assume there is no shadowing in the uplink. The other configuration parameters are listed in Table 11.

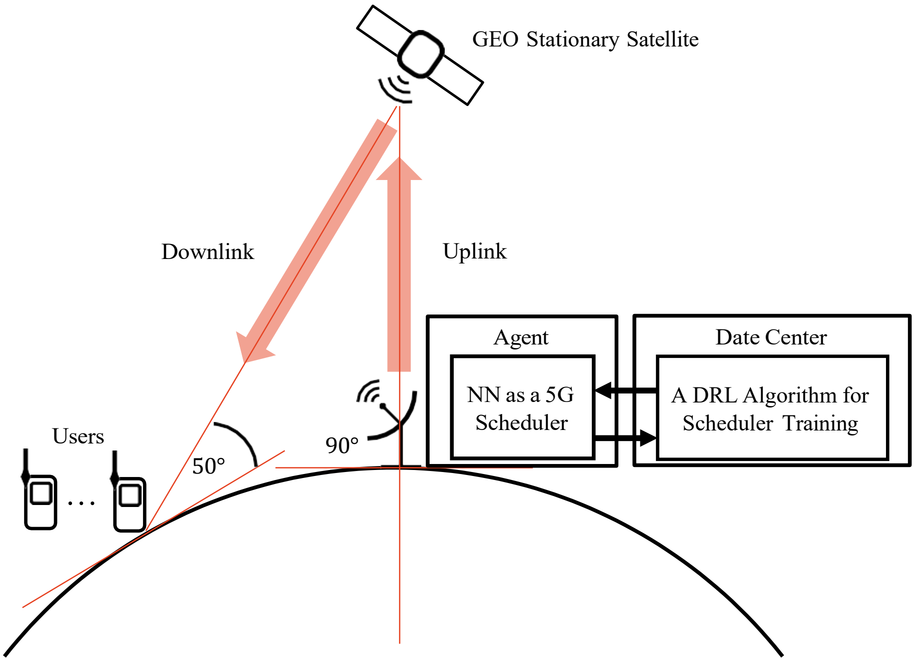


Figure 54: Setup of a forward link of 5G NTN. Table 11:

Simulation Setup [175]

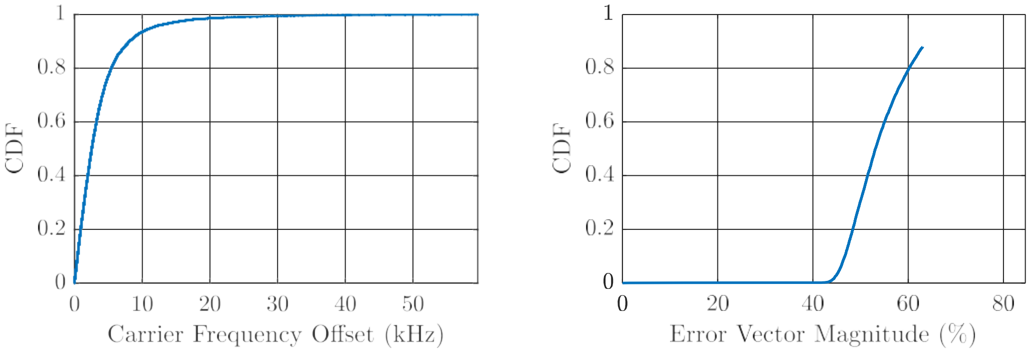
	Uplink	Downlink
Effective isotropic radiated Power	91.2 dBm	76.2 dBm
Antenna gain-to-noise-temperature	15.9 dB	28.0 dB
Bandwidth	144 MHz	144 MHz
Subcarrier spacing	120 KHz	120 KHz
Carrier frequency	20 GHz	30 GHz

13.1.2 Evaluation of the physical layer of 5G NTN

We first evaluate the physical layer performance of 5G NTN in the forward link in MAT- LAB. Specifically, we transmit 10^4 synchronization signal (SS) bursts over this forward link. Then, we measure the carrier frequency offset (CFO) and the error vector magni- tude (EVM) of the bursts received at the user, as shown in Fig. 55. The average CFO is measured as 3.84 KHz, which is far less than the subcarrier spacing (e.g., 120 KHz). We measure the average EVM of the SS burst as 54.36 %, which leads to 1.40 % block error rate of the broadcast channel transmitted in the SS Bursts. These results indicate the feasibility of transmissions of the 5G waveform over NTNs.

13.1.3 Evaluation of AI-enabled software-defined satellite architecture 5G NTN scheduler design

In this part, we evaluate a software-defined satellite architecture that uses deep reinforce- ment learning (DRL) to design the scheduler for a 5G NTN [184], as shown in Fig. 54. In this architecture, a DRL algorithm, namely deep deterministic policy gradient (DDPG), is used to train an NN as a scheduler. Here, the scheduler takes the queue state infor- mation (QSI) and the channel state information (CSI) as its input and outputs the users to be scheduled. The reward function is defined as the packet loss probabilities, which is evaluated numerically according to the CSI reported by the users via a return link.



(a) Carrier frequency offset of SS bursts received at the user.

(b) Error vector magnitude of SS bursts received at the user.

Figure 55: The performance of SS bursts over the forward link.

Based on the state (i.e., QSI, CSI), action (the scheduled users), and the reward (packet loss probabilities) in the current transmission time interval (TTI), as well as the state in the next TTI, DDPG keeps updating the weights and biases of the scheduler in order to increase the long-term reward of each user.

In reality, the CSI is delayed due to long communication distances between the users, the GEO satellite, and the ground station. To illustrate the impact caused by the delayed CSI, we will compare two cases: 1) a realistic case where the CSI is delayed, and 2) an ideal case where the CSI is not delayed and is directly available at the ground station. The packet size is 150 bytes and the average packet arrival rate is 10^3 packets/second.

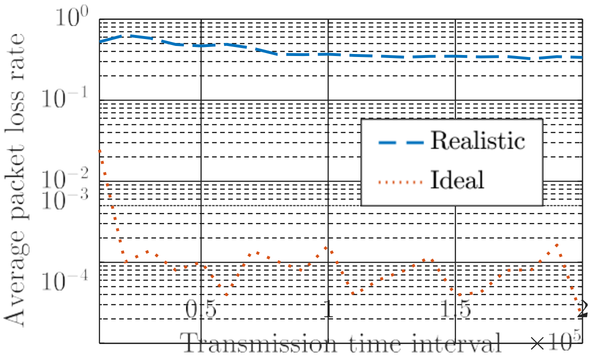


Figure 56: Average packet loss probability of 5 users.

The average packet loss probabilities of 5 users over 10^5 TTIs are shown in Fig. 56. The legends of the realistic case with the delayed CSI and the ideal case with perfect CSI are “Realistic” and“Ideal”, respectively. For both cases, average packet loss rates decrease with time. However, the ideal case has a much lower packet loss rate compared with the realistic case. This is because the DDPG algorithm determines the action based on the delayed CSI in the realistic case, which leads to non-optimal scheduling decisions. In future, methods to estimate (or predict) the CSI could be investigated to address this issue. Alternatively, we can optimised the scheduling policy based on the QSI and the large-scale channel gains that vary slowly.

13.2 Conclusion and Future Directions

In this section, we have proposed the AI-enabled SDN architecture for 5G NTNs. We have identified the possible research problems to design and develop the architecture. Also, we have performed the preliminary evaluation of the proposed architecture. Specifically, the physical layer simulation of the forward link shows the feasibility of the transmission of the 5G waveform over satellite networks. Further, we developed and evaluated the deep reinforcement learning algorithm for the scheduler design in the proposed architecture, in which we find that the delay CSI in 5G NTNs leads to low reliability of the transmissions. In future, we could keep tracking the 3GPP standardisation process of 5G NTNs. Follow- ing the standardisation of 5G NTNs, we can design the SDN architecture to enable the programmability of 5G NTNs for different SDN applications. Then, AI algorithms can be developed to optimise the end-to-end performance over the 5G NTNs based on the SDN architectures.

References

[1] B. R. Vojcic, L. B. Milstein, and R. L. Pickholtz, “Downlink DS CDMA performance over a mobile satellite channel,” *IEEE transactions on vehicular technology*, vol. 45, no. 3, pp. 551–560, 1996.

[2] N. Letzepis and A. J. Grant, “Capacity of the multiple spot beam satellite channel with Rician fading,” *IEEE transactions on information theory*, vol. 54, no. 11, pp. 5210–5222, 2008.

[3] I. Ali, N. Al-Dhahir, and J. E. Hershey, “Doppler characterization for LEO satellites,” *IEEE transactions on communications*, vol. 46, no. 3, pp. 309–313, 1998.

[4] J. Lin, Z. Hou, Y. Zhou, L. Tian, and J. Shi, “Map estimation based on doppler characterization in broadband and mobile LEO satellite communications,” in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, IEEE, 2016.

[5] J. Li, W. Xiong, G. Sun, Z. Wang, Y. Huang, and M. Shen, “Doppler-robust high-spectrum-efficiency VCM-OFDM scheme for low Earth orbit satellites broadband data transmission,” *IET Communications*, vol. 12, no. 1, pp. 35–43, 2017.

[6] I. Ali, N. A.-Dhahir, J. E. Hershey, G. J. Saulnier, and R. Nelson, “Doppler as a new dimension for multiple access in LEO satellite systems,” *Int. J. Satell. Commun.*, vol. 15, pp. 269–279, Dec. 1998.

[7] F. Babich, G. Lombardi, and E. Valentinuzzi, “Variable order Markov modelling for LEO mobile satellite channels,” *Electronics Letters*, vol. 35, no. 8, pp. 621–623, 1999.

[8] J. Lei and M. A. Vazquez-Castro, “Joint power and carrier allocation for the multi-beam satellite downlink with individual SINR constraints,” in *2010 IEEE International Conference on Communications*, pp. 1–5, IEEE, 2010.

[9] Y. Guan, F. Geng, and J. H. Saleh, “Review of high throughput satellites: Market disruptions, affordability-throughput map, and the cost per bit/second decision tree,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 5, pp. 64–80, 2019.

[10] “Sky mesh spot beam map.” https://www.skymesh.net.au/support/troubleshooting/sky-muster-spot-beams/. [Online; accessed 14-August-2020].

[11] Y. Couble, C. Rosenberg, E. Chaput, J.-B. Dupé, C. Baudoin, and A.-L. Beylot, “Two-color scheme for a multi-beam satellite return link: Impact of interference coordination,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 5, pp. 993–1003, 2018.

[12] W. Zheng, B. Li, J. Chen, and J. Wu, “A novel dual-size interleaved spot-beam ar- chitecture for mobile satellite communications,” *2013 15th International Conference on Advanced Communications Technology (ICACT)*, pp. 794–798, 2013.

[13] Z. Li, H. Song, Q. Cui, and B. Liu, “A novel multi-beam architecture in mobile satellite communications,” in *2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, 2014.

[14] Y. Couble, E. Chaput, J.-B. Dupé, C. Bes, T. Deleu, C. Baudoin, and A.-L. Beylot, "Performance evaluation of aggressive frequency reuse schemes in the return link of multibeam satellites," 2018.

[15] H. Fenech, S. Amos, A. Tomatis, and V. Soumpholphakdy, "High throughput satellite systems: An analytical approach," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 1, pp. 192–202, 2015.

[16] S. Rao, M. Tang, and C.-C. Hsu, "Multiple beam antenna technology for satellite communications payloads," *Applied Computational Electromagnetics Society Journal*, vol. 21, pp. 353–363, 11 2006.

[17] A. Jacomb-Hood and E. Lier, "Multibeam active phased arrays for communications satellites," *IEEE Microwave Magazine*, vol. 1, no. 4, pp. 40–47, 2000.

[18] M. Hasan and C. Bianchi, "Ka band enabling technologies for high throughput satellite (HTS) communications," *International Journal of Satellite Communications and Networking*, vol. 34, no. 4, pp. 483–501, 2016.

[19] H.-T. Zhang, W. Wang, M.-P. Jin, and X.-P. Lu, "An active phased array antenna for broadband mobile satellite communications at Ka-band," in *2016 CIE International Conference on Radar (RADAR)*, pp. 1–3, IEEE, 2016.

[20] D. SERRANO-VELARDE, E. Lance, H. Fenech, and G. E. Rodriguez-guisantes, "Novel dimensioning method for high-throughput satellites: forward link," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 3, pp. 2146–2163, 2014.

[21] A. Kyrgiazos, B. Evans, P. Thompson, P. T. Mathiopoulos, and S. Papaharalabos, "A terabit/second satellite system for European broadband access: a feasibility study," *International Journal of Satellite Communications and Networking*, vol. 32, no. 2, pp. 63–92, 2014.

[22] M. A. Vazquez, A. Perez-Neira, D. Christopoulos, S. Chatzinotas, B. Ottersten, P.-D. Arapoglou, A. Ginesi, and G. Tarocco, "Precoding in multibeam satellite communications: Present and future challenges," *IEEE Wireless Communications*, vol. 23, no. 6, pp. 88–95, 2016.

[23] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Frame based precoding in satellite communications: A multicast approach," in *2014 7th Advanced Satellite Multimedia Systems Conference and the 13th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, pp. 293–299, IEEE, 2014.

[24] G. Taricco, "Linear precoding methods for multi-beam broadband satellite systems," in *European Wireless 2014; 20th European Wireless Conference*, pp. 1–6, VDE, 2014.

[25] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Multicast multigroup precoding and user scheduling for frame-based satellite communications," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4695–4707, 2015.

[26] V. Joroughi, M. A. Vázquez, and A. I. Pérez-Neira, “Generalized multicast multibeam precoding for satellite communications,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 952–966, 2017.

[27] A. Bandi, S. Chatzinotas, B. Ottersten, *et al.*, “Joint scheduling and precoding for frame-based multigroup multicasting in satellite communications,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2019.

[28] A. I. Perez-Neira, M. A. Vazquez, M. B. Shankar, S. Maleki, and S. Chatzinotas, “Signal processing for high-throughput satellites: Challenges in new interference-limited scenarios,” *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 112–131, 2019.

[29] Y. C. Silva and A. Klein, “Linear transmit beamforming techniques for the multigroup multicast scenario,” *IEEE Transactions on Vehicular Technology*, vol. 58, no. 8, pp. 4353–4367, 2009.

[30] D. Christopoulos, S. Chatzinotas, and B. Ottersten, “Weighted fair multicast multigroup beamforming under per-antenna power constraints,” *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5132–5142, 2014.

[31] B. Devillers, A. Pérez-Neira, and C. Mosquera, “Joint linear precoding and beamforming for the forward link of multi-beam broadband satellite systems,” in *2011 IEEE Global Telecommunications Conference-GLOBECOM 2011*, pp. 1–6, IEEE, 2011.

[32] T. Yoo and A. Goldsmith, “On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming,” *IEEE Journal on selected areas in communications*, vol. 24, no. 3, pp. 528–541, 2006.

[33] A. Guidotti and A. Vanelli-Coralli in *2018 9th Advanced Satellite Multimedia Systems Conference and the 15th Signal Processing for Space Communications Workshop (ASMS/SPSC)*.

[34] M. A. Vázquez, M. B. Shankar, C. I. Kourogiorgas, P.-D. Arapoglou, V. Icolari, S. Chatzinotas, A. D. Panagopoulos, and A. I. Pérez-Neira, “Precoding, schedul- ing, and link adaptation in mobile interactive multibeam satellite systems,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 5, pp. 971–980, 2018.

[35] A. Gharanjik, B. S. MR, P.-D. Arapoglou, M. Bengtsson, and B. Ottersten, “Robust precoding design for multibeam downlink satellite channel with phase uncertainty,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3083–3087, IEEE, 2015.

[36] L. You, A. Liu, W. Wang, and X. Gao, “Outage constrained robust multigroup multicast beamforming for multi-beam satellite communication systems,” *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 352–355, 2018.

[37] Y. Yan, W. Yang, B. Zhang, D. Guo, and G. Ding, “Outage constrained robust beamforming for sum rate maximization in multi-beam satellite systems,” *IEEE Communications Letters*, vol. 24, no. 1, pp. 164–168, 2020.

[38] A. Gharanjik, B. S. MR, P.-D. Arapoglou, and B. Ottersten, “Multiple gateway transmit diversity in Q/V band feeder links,” *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 916–926, 2014.

[39] T. De Cola and M. Mongelli, “Adaptive time window linear regression for outage prediction in Q/V band satellite systems,” *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 808–811, 2018.

[40] B. Roy, S. Poulenard, S. Dimitrov, R. Barrios, D. Giggenbach, A. Le Kernec, and M. Sotom, “Optical feeder links for high throughput satellites,” in *2015 IEEE International Conference on Space Optical Systems and Applications (ICSOS)*, pp. 1–6, IEEE, 2015.

[41] E. Zedini, A. Kammoun, and M.-S. Alouini, “Performance of Multibeam Very High Throughput Satellite Systems Based on FSO Feeder Links with HPA Nonlinearity,” *IEEE Transactions on Wireless Communications*, 2020.

[42] G. Zheng, S. Chatzinotas, and B. Ottersten, “Multi-gateway cooperation in multibeam satellite systems,” in *2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications-(PIMRC)*, pp. 1360–1364, IEEE, 2012.

[43] D. Christopoulos, H. Pennanen, S. Chatzinotas, and B. Ottersten, “Multicast multigroup precoding for frame-based multi-gateway satellite communications,” in *2016 8th Advanced Satellite Multimedia Systems Conference and the 14th Signal Processing for Space Communications Workshop (ASMS/ SPSC)*, pp. 1–6, IEEE, 2016.

[44] V. Joroughi, M. Á . Vázquez, and A. I. Pérez-Neira, “Precoding in multigateway multibeam satellite systems,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4944–4956, 2016.

[45] N. Song, T. Yang, and M. Haardt, “Efficient hybrid space-ground precoding techniques for multi-beam satellite systems,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6284–6288, IEEE, 2017.

[46] V. Joroughi, M. Á . Vázquez, A. I. Pérez-Neira, and B. Devillers, “Onboard beam generation for multibeam satellite systems,” *IEEE transactions on wireless communications*, vol. 16, no. 6, pp. 3714–3726, 2017.

[47] C. Henry, “ViaSat plans massive ground network of smaller gateways for ViaSat-2 and ViaSat-3 satellites.” https://spacenews.com/viasat-plans-massive-ground- network-of-smaller-gateways-for-viasat-2-and-viasat-3-satellites/.

[48] I. Del Portillo, B. G. Cameron, and E. F. Crawley, “A technical comparison of three low earth orbit satellite constellation systems to provide global broadband,” *Acta Astronautica*, vol. 159, pp. 123–135, 2019.

[49] T. Rossi, M. De Sanctis, F. Maggio, M. Ruggieri, C. Hibberd, and C. Togni, “Smart gateway diversity optimization for ehf satellite networks,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 1, pp. 130–141, 2019.

[50] G. Hill, “Satellite internet gateway location whitepaper.” https://x2n.com/blog/satellite-internet-gateway-location-whitepaper/.

[51] K. Yang, B. Zhang, and D. Guo, “Partition-based joint placement of gateway and controller in sdn-enabled integrated satellite-terrestrial networks,” *Sensors*, vol. 19, no. 12, p. 2774, 2019.

[52] J. Liu, Y. Shi, L. Zhao, Y. Cao, W. Sun, and N. Kato, “Joint placement of controllers and gateways in sdn-enabled 5g-satellite integrated network,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 2, pp. 221–232, 2018.

[53] N. Torkzaban, A. Gholami, J. S. Baras, and C. Papagianni, “Joint satellite gateway placement and routing for integrated satellite-terrestrial networks,” in *ICC 2020- 2020 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2020.

[54] Q. Chen, L. Yang, X. Liu, J. Guo, S. Wu, and X. Chen, “Multiple gateway placement in large-scale constellation networks with inter-satellite links,” *International Journal of Satellite Communications and Networking*.

[55] J. P. Choi and V. W. Chan, “Optimum power and beam allocation based on traffic demands and channel conditions over satellite downlinks,” *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2983–2993, 2005.

[56] X. Alberti, J. Cebrian, A. Del Bianco, Z. Katona, J. Lei, M. Vazquez-Castro, A. Zanus, L. Gilbert, and N. Alagha, “System capacity optimization in time and frequency for multibeam multi-media satellite systems,” in *2010 5th Advanced Satellite Multimedia Systems Conference and the 11th Signal Processing for Space Communications Workshop*, pp. 226–233, IEEE, 2010.

[57] F. Qi, L. Guangxia, F. Shaodong, and G. Qian, “Optimum power allocation based on traffic demand for multi-beam satellite communication systems,” in *2011 IEEE 13th International Conference on Communication Technology*, pp. 873–876, IEEE, 2011.

[58] A. I. Aravanis, B. S. MR, P.-D. Arapoglou, G. Danoy, P. G. Cottis, and B. Ottersten, “Power allocation in multibeam satellite systems: A two-stage multi-objective optimization,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3171–3182, 2015.

[59] G. Cocco, T. De Cola, M. Angelone, Z. Katona, and S. Erl, “Radio resource man- agement optimization of flexible satellite payloads for dvb-s2 systems,” *IEEE Trans- actions on Broadcasting*, vol. 64, no. 2, pp. 266–280, 2018.

[60] M. Jia, X. Zhang, X. Gu, Q. Guo, Y. Li, and P. Lin, “Interbeam interference constrained resource allocation for shared spectrum multibeam satellite communication systems,” *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6052–6059, 2019.

[61] S. Chatzinotas, G. Zheng, and B. Ottersten, “Joint precoding with flexible power constraints in multibeam satellite systems,” in *2011 IEEE Global Telecommunications Conference-GLOBECOM 2011*, pp. 1–5, IEEE, 2011.

[62] G. Zheng, S. Chatzinotas, and B. Ottersten, “Generic optimization of linear precoding in multibeam satellite systems,” *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2308–2320, 2012.

[63] Z. Luo, D. Yang, H. Wang, and J. Kuang, "Weighted fair precoding based on traffic demands for multibeam satellite systems," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, pp. 1–5, IEEE, 2019.

[64] ViaSat and F. C. Commission, "Attachment A technical information to supplement schedule S." http://licensing.fcc.gov/myibfs/download.do?attachment$_{k}ey$ = 1093081.

[65] C. Miller, "ViaSat's Global Ka-Band Constellation and Commercial SATCOM Applicability for Australia." https://static1.squarespace.com/static/5274112ae4b02d3f058d4348/t/5a14f1e3085229dcccf1c987/1511322093111/2017-2-8b.pdf.

[66] M. Torrieri, "VHTS: Soaring to Unprecedented Heights." http://interactive.satellitetoday.com/via/january-2020/vhts-soaring-to-unprecedented-heights/.

[67] P. Haines, "Current satcom trends." www.joanneum.at/fileadmin/UNTERNEHMEN/news/Zukunftskonferenz_ 2018/ZK_ 18_DIG_Philip_Haines.pdf.

[68] L. SnT, University of Luxembourg, "LiveSatPreDem – Live Satellite Precoding Demonstration." https://wwwen.uni.lu/snt/research/sigcom/projects/ pre- dem prec_oding dem_onstrator for b_roa_dband syst_em forw_ard links_.

[69] J. C. Merlano Duncan, J. Querol Borras, N. Maturo, J. Krivochiza, D. Spano, S. Norshahida, L. Martinez Marrero, S. Chatzinotas, and B. Ottersten, "Hardware Precoding Demonstration in Multi-Beam UHTS Communications under Realistic Payload Characteristics," in *Proceedings of the 37th International Communications Satellite Systems Conference*, 2019.

[70] N. Maturo, J. C. M. Duncan, J. Krivochiza, J. Querol, D. Spano, S. Chatzinotas, and B. Ottersten, "Demonstrator of precoding technique for a multi-beams satellite system," in *2019 8th International Workshop on Tracking, Telemetry and Command Systems for Space Applications (TTC)*, pp. 1–8, IEEE, 2019.

[71] J. Duncan, J. Krivochiza, S. Andrenacci, S. Chatzinotas, and B. Ottersten, "Hard- ware demonstration of precoded communications in multi-beam uhts systems," 2018.

[72] L. Lei, E. Lagunas, Y. Yuan, M. G. Kibria, S. Chatzinotas, and B. Ottersten, "Beam illumination pattern design in satellite networks: Learning and optimization for efficient beam hopping," *IEEE Access*, vol. 8, 2020.

[73] V. Joroughi, E. Lagunas, S. Andrenacci, N. Maturo, S. Chatzinotas, J. Grotz, and B. Ottersten, "Deploying joint beam hopping and precoding in multibeam satellite networks with time variant traffic," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1081–1085, IEEE, 2018.

[74] M. G. Kibria, E. Lagunas, N. Maturo, D. Spano, and S. Chatzinotas, "Precoded cluster hopping in multi-beam high throughput satellite systems," in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2019.

[75] S. Dimitrov, S. Erl, S. Jaeckel, J. Rodriguez, A. Yun, A. Kyrgiazos, B. Evans, O. Vidal, and P. Inigo, "Radio resource management for forward and return links in high throughput satellite systems," in *Proceedings of 20th Ka and Broadband Communications Conference, Vietri sul Mare/Salerno, Italy*, pp. 1–3, 2014.

[76] Y. Couble, E. Chaput, T. Deleu, C. Baudoin, J.-B. Dupé, C. Bés, and A.-L. Beylot, "Interference-aware frame optimization for the return link of a multi-beam satellite," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2017.

[77] V. Boussemart, M. Berioli, F. Rossetto, and M. Joham, "On the achievable rates for the return-link of multi-beam satellite systems using successive interference cancellation," pp. 217–223, 2011.

[78] V. Boussemart, M. Berioli, and F. Rossetto, "User scheduling for large multi-beam satellite MIMO systems," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 1800–1804, IEEE, 2011.

[79] U. Y. Ng, A. Kyrgiazos, and B. Evans, "Interference coordination for the return link of a multibeam satellite system," in *2014 7th Advanced Satellite Multimedia Systems Conference and the 13th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, pp. 366–373, 2014.

[80] J. R. Bejarano, C. Miguel Nieto, and F. J. Ruiz Piñar, "MF-TDMA Scheduling Algorithm for Multi-Spot Beam Satellite Systems Based on Co-Channel Interference Evaluation," *IEEE Access*, vol. 7, pp. 4391–4399, 2019.

[81] B. Makki, T. Svensson, G. Cocco, T. de Cola, and S. Erl, "On the throughput of the return-link multi-beam satellite systems using genetic algorithm-based schedulers," in *2015 IEEE International Conference on Communications (ICC)*, pp. 838–843, IEEE, 2015.

[82] A. Kyrgiazos, P. Thompson, and B. Evans, "Gateway diversity via flexible resource allocation in a multibeam SS-TDMA system," *IEEE communications letters*, vol. 17, no. 9, pp. 1762–1765, 2013.

[83] A. Kyrgiazos, B. G. Evans, and P. Thompson, "On the gateway diversity for high throughput broadband satellite systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5411–5426, 2014.

[84] M. Marchese, A. Moheddine, F. Patrone, T. d. Cola, and M. Mongelli, "QoS-Aware Handover Strategies for Q/V Feeder Links in VHTS Systems," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pp. 1–7, 2020.

[85] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: Enabling innovation in campus networks," *SIGCOMM Comput. Commun. Rev.*, vol. 38, p. 69–74, Mar. 2008.

[86] H. Wang, Z. Liu, Z. Cheng, Y. Miao, W. Feng, and N. Ge, "Maximization of link capacity by joint power and spectrum allocation for smart satellite transponder," in *2017 23rd Asia-Pacific Conference on Communications (APCC)*, pp. 1–6, 2017.

[87] H. Al-Hraishawi, N. Maturo, E. Lagunas, and S. Chatzinotas, “Perceptive packet scheduling for carrier aggregation in satellite communication systems,” in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2020.

[88] E. Lutz, M. Werner, and A. Jahn, *Satellite systems for personal and broadband communications*. Springer Science & Business Media, 2012.

[89] H. Wang, A. Liu, X. Pan, and J. Yang, “Optimization of power allocation for multiusers in multi-spot-beam satellite communication systems,” *Mathematical Problems in engineering*, vol. 2014, 2014.

[90] U. Park, H. W. Kim, D. S. Oh, and B. J. Ku, “Flexible bandwidth allocation scheme based on traffic demands and channel conditions for multi-beam satellite systems,” in *2012 IEEE Vehicular Technology Conference (VTC Fall)*, pp. 1–5, IEEE, 2012.

[91] M. Jia, X. Zhang, X. Gu, Q. Guo, Y. Li, and P. Lin, “Interbeam interference constrained resource allocation for shared spectrum multibeam satellite communication systems,” *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6052–6059, 2018.

[92] G. Cocco, T. De Cola, M. Angelone, Z. Katona, and S. Erl, “Radio resource manage- ment optimization of flexible satellite payloads for DVB-S2 systems,” *IEEE Trans- actions on Broadcasting*, vol. 64, no. 2, pp. 266–280, 2017.

[93] J. P. Choi and V. W. Chan, “Optimum power and beam allocation based on traffic demands and channel conditions over satellite downlinks,” *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2983–2993, 2005.

[94] R. Chen, X. Hu, X. Li, and W. Wang, “Optimum power allocation based on traffic matching service for multi-beam satellite system,” in *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, pp. 655–659, IEEE, 2020.

[95] C. Daehnick, I. Klinghoffer, B. Maritz, and B. Wiseman, “Large leo satellite constellations: Will it be diffrent this time?,” tech. rep., McKinsey & Company. [Available Online]: https://www.mckinsey.com/industries/aerospace-and-defense/our-insights/large-leo-satellite-constellations-will-it-be-different-this-time, May 2020.

[96] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, “Broadband LEO satellite com- munications: Architectures and key technologies,” *IEEE Wireless Communications*, vol. 26, no. 2, pp. 55–61, 2019.

[97] M. Harris, “Tech giants race to build orbital internet,” *IEEE Spectrum*, vol. 55, pp. 123–135, Jun. 2018.

[98] M. Gurman, “Apple has secret team working on satellites to beam data to devices,” tech. rep., Bloomberg Technology News. [Available Online]: https://www.bloomberg.com/news/articles/2019-12-20/apple-has-top-secret- team-working-on-internet-satellites, Dec. 2019.

[99] C. Wang, D. Bian, S. Shi, J. Xu, and G. Zhang, “A Novel Cognitive Satellite Network With GEO and LEO Broadband Systems in the Downlink Case,” *IEEE Access*, vol. 6, pp. 25987–26000, 2018.

[100] 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, pp. 127–133, Aug. 2017.

[101] T. Li, H. Zhou, H. Luo, and S. Yu, “SERvICE: A software defined framework for integrated space-terrestrial satellite communication,” *IEEE Trans. Mobile Computing*, vol. 17, pp. 703–716, Mar. 2018.

[102] S. G. Glisic, J. J. Talvitie, T. Kumpumaki, M. Latva-aho, J. H. Iinatti, and T. J. Poutanen, “Design study for a CDMA-based LEO satellite network: downlink system level parameters,” *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 9, pp. 1796–1808, 1996.

[103] U. A. Waheed and K. D. Vimal, “Downlink performance of multi-beam multi-satellite CDMA-based LEO satellite system with power control,” in *IEEE Global Telecommunications Conference Globecom’00*, vol. 2, pp. 1140–1144, IEEE, 200.

[104] H. Fu, G. Bi, and K. Arichandran, “Capacity enhancement with adaptive arrays in a CDMA-based LEO satellite system,” in *IEEE 51st Vehicular Technology Conference Proceedings VTC’00-Spring*, vol. 3, pp. 1979–1982, IEEE, 200.

[105] C. Tarver, A. Balatsoukas-Stimming, and J. R. Cavallaro, “Design and implementation of a neural network based predistorter for enhanced mobile broadband,” in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, pp. 296–301, IEEE, 2019.

[106] T. Ishiguro, T. Hara, and M. Okada, “Post-Compensation Technique for Carrier Superposed Satellite Channel Including Nonlinear TWTA,” *IEICE transactions on communications*, vol. 95, no. 11, pp. 3420–3427, 2012.

[107] J. Malone and M. A. Wickert, “Practical Volterra equalizers for wideband satellite communications with TWTA nonlinearities,” in *2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, pp. 48–53, IEEE, 2011.

[108] S. Jung, J. G. Ryu, D.-G. Oh, and H. Yu, “Low-complexity nonlinearity post compensator for shared band transmission in satellite communication,” in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 1–6, IEEE, 2018.

[109] J. K. Cavers, “Amplifier linearization using a digital predistorter with fast adaptation and low memory requirements,” *IEEE transactions on vehicular technology*, vol. 39, no. 4, pp. 374–382, 1990.

[110] C. Eun and E. J. Powers, “A new Volterra predistorter based on the indirect learning architecture,” *IEEE transactions on signal processing*, vol. 45, no. 1, pp. 223–227, 1997.

[111] J. Peroulas, “Digital predistortion using machine learning algorithms,” 2016. Accessed: 2020-08-08.

[112] J. Peng, S. He, B. Wang, Z. Dai, and J. Pang, “Digital predistortion for power amplifier based on sparse Bayesian learning,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 9, pp. 828–832, 2016.

[113] Z. Wang, Y. Wang, C. Song, T. Chen, and W. Cheng, “Deep neural nets based power amplifier non-linear pre-distortion,” *JPhCS*, vol. 887, no. 1, p. 012049, 2017.

[114] D. Wang, M. Zhang, Z. Li, Y. Cui, J. Liu, Y. Yang, and H. Wang, “Nonlinear decision boundary created by a machine learning-based classifier to mitigate nonlinear phase noise,” in *2015 European Conference on Optical Communication (ECOC)*, pp. 1–3, IEEE, 2015.

[115] D. Wang, M. Zhang, M. Fu, Z. Cai, Z. Li, H. Han, Y. Cui, and B. Luo, “Nonlinearity mitigation using a machine learning detector based on k-nearest neighbors,” *IEEE Photonics Technology Letters*, vol. 28, no. 19, pp. 2102–2105, 2016.

[116] T. Delamotte, K. Storek, and A. Knopp, “MIMO Processing for Satellites in the 5G Era,” in *2019 IEEE 2nd 5G World Forum (5GWF)*, pp. 629–635, 2019.

[117] L. G. Barbero and J. S. Thompson, “A Fixed-Complexity MIMO Detector Based on the Complex Sphere Decoder,” in *2006 IEEE 7th Workshop on Signal Processing Advances in Wireless Communications*, pp. 1–5, 2006.

[118] L. Fang, L. Xu, and D. D. Huang, “Low Complexity Iterative MMSE-PIC Detection for Medium-Size Massive MIMO,” *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 108–111, 2016.

[119] M. Berceanu, C. Voicu, and S. Halunga, “The performance of an uplink Large Scale MIMO system with MMSE-SIC detector,” in *2019 International Conference on Military Communications and Information Systems (ICMCIS)*, pp. 1–4, 2019.

[120] B. Vucetic and J. Yuan, *Space-time coding*. John Wiley and Sons, 2003.

[121] N. Aboutorab, W. Hardjawana, and B. Vucetic, “A New Iterative Doppler-Assisted Channel Estimation Joint With Parallel ICI Cancellation for High-Mobility MIMO-OFDM Systems,” *IEEE Transactions on Vehicular Technology*, vol. 61, no. 4,
pp. 1577–1589, 2012.

[122] N. Aboutorab, W. Hardjawana, and B. Vucetic, “Application of compressive sensing to channel estimation of high mobility OFDM systems,” in *2013 IEEE International Conference on Communications (ICC)*, pp. 4946–4950, 2013.

[123] T. Minka, “Expectation Propagation for Approximate Bayesian Inference,” vol. 17, 01 2001.

[124] J. Céspedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, “Expectation Propagation Detection for High-Order High-Dimensional MIMO Systems,” *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 2840–2849, 2014.

[125] K. Takeuchi and C. Wen, “Rigorous dynamics of expectation-propagation signal detection via the conjugate gradient method,” in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, 2017.

[126] G. Yao, G. Yang, J. Hu, and C. Fei, “A Low Complexity Expectation Propagation Detection for Massive MIMO System,” in *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2018.

[127] H. Wang, A. Kosasih, C. Wen, S. Jin, and W. Hardjawana, “Expectation Propagation Detector for Extra-Large Scale Massive MIMO,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2036–2051, 2020.

[128] C. Jeon, R. Ghods, A. Maleki, and C. Studer, “Optimal data detection in large MIMO,” *CoRR*, vol. abs/1811.01917, 2018.

[129] A. Kosasih, W. Hardjawana, B. Vucetic, and C. Wen, “A Linear Bayesian Learning Receiver Scheme for Massive MIMO Systems,” in *2020 IEEE Wireless Communi- cations and Networking Conference (WCNC)*, pp. 1–6, 2020.

[130] W. Guo, “Explainable Artificial Intelligence for 6G: Improving Trust between Human and Machine,” *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39–45, 2020.

[131] J. Zhang, T. Q. Duong, A. Marshall, and R. Woods, “Key generation from wireless channels: A review,” *IEEE Access*, vol. 4, pp. 614–626, Jan. 2016.

[132] J. W. Wallace and R. K. Sharma, “Automatic secret keys from reciprocal MIMO wireless channels: Measurement and analysis,” *IEEE Trans. Inf. Forensics Security*, vol. 5, pp. 381–392, Sep. 2010.

[133] Y. Peng, P. Wang, W. Xiang, and Y. Li, “Secret key generation based on estimated channel state information for TDD-OFDM systems over fading channels,” *IEEE Trans. Wireless Commun.*, vol. 16, pp. 5176–5186, Aug. 2017.

[134] H. Liu, J. Yang, Y. Wang, Y. Chen, and C. E. Koksal, “Group secret key generation via received signal strength: Protocols, achievable rates and implementation,” *IEEE Trans. Mobile Comput.*, vol. 13, pp. 2820–2835, Dec. 2014.

[135] C. D. T. Thai, J. Lee, and T. Q. S. Quek, “Physical-layer secret key generation with colluding untrusted relays,” *IEEE Trans. Wireless Commun.*, vol. 15, pp. 1517– 1530, Feb. 2016.

[136] Z. Ji, Y. Zhang, Z. He, K. Lin, B. Li, P. L. Yeoh, and H. Yin, “Vulnerabilities of physical layer secret key generation against environment reconstruction based attacks,” *IEEE Wireless Commun. Lett.*, vol. 9, pp. 693–697, May 2020.

[137] P. Azmi, M. Forouzesh, A. Kuhestani, and P. L. Yeoh, “Covert communication and secure transmission over untrusted relaying networks in the presence of multiple wardens,” *IEEE Trans. Commun.*, vol. 68, pp. 3737–3749, Mar. 2020.

[138] A. Kuhestani, A. Mohammadi, and P. L. Yeoh, “Optimal power allocation and secrecy sum rate in two-way untrusted relaying networks using a friendly jammer,” *IEEE Trans. Commun.*, vol. 66, pp. 2671–2684, Feb. 2018.

[139] A. Kuhestani, A. Mohammadi, K.-K. Wong, P. L. Yeoh, M. Moradikia, and M. R. A. Khandaker, “Optimal power allocation by imperfect hardware analysis in untrusted relaying networks,” *IEEE Trans. Wireless Commun.*, vol. 17, pp. 4302–4314, Jul. 2018.

[140] J. Vilela, M. Bloch, J. Barros, and S. W. McLaughlin, “Wireless secrecy regions with friendly jamming,” *IEEE Trans. Inf. Forensics Secur.*, vol. 6, pp. 256–266, Jun. 2011.

[141] X. Ding, T. Song, Y. Zou, and X. Chen, “Security-reliability tradeoff for friendly jammer assisted user-pair selection in the face of multiple eavesdroppers,” *IEEE Access*, vol. 4, pp. 8386–8393, 2016.

[142] S. Bayat, R. H. Y. Louie, Z. Han, B. Vucetic, and Y. Li, “Physical-layer security in distributed wireless networks using matching theory,” *IEEE Trans. Inf. Forensics Secur.*, vol. 8, pp. 717–732, May 2013.

[143] S. Yan, N. Yang, I. Land, R. Malaney, and J. Yuan, “Three artificial-noise aided se- cure transmission schemes in wiretap channels,” *IEEE Trans.Veh. Technol.*, vol. 67, pp. 3669–3673, Apr. 2018.

[144] A. Al-Hourani, S. Kandeepan, and S. Lardner, “Optimal lap altitude for maximum coverage,” *IEEE Wireless Commun. Lett.*, vol. 3, pp. 569–572, Dec. 2014.

[145] M. Alzenad, A. El-Keyi, and H. Yanikomeroglu, “3-D placement of an unmanned aerial vehicle base station for maximum coverage of users with different QoS requirements,” *IEEE Wireless Commun. Lett.*, vol. 7, pp. 38–41, Feb. 2018.

[146] S. Zhang, H. Zhang, Q. He, K. Bian, and L. Song, “Joint trajectory and power optimization for UAV relay networks,” *IEEE Commun. Lett.*, vol. 22, pp. 161–164, Jan. 2018.

[147] Y. Zhou, P. L. Yeoh, H. Chen, Y. Li, R. Schober, L. Zhuo, and B. Vucetic, “Improving physical layer security via a uav friendly jammer for unknown eavesdropper location,” *IEEE Trans. Vehic. Technol.*, vol. 67, pp. 11280–11284, Nov. 2018.

[148] Y. Zhou, C. Pan, P. L. Yeoh, K. Wang, M. Elkashlan, B. Vucetic, and Y. Li, “Secure Communications for UAV-Enabled Mobile Edge Computing System,” *IEEE Trans. Commun.*, vol. 68, pp. 376–388, Jan. 2020.

[149] 3GPP, “Study on Architecture Aspects for Using Satellite Access in 5G,” SP SP- 181253, 3GPP, 2019.

[150] O. Kodheli, E. Lagunas, N. Maturo, S. K. Sharma, B. Shankar, J. Montoya, J. Dun- can, D. Spano, S. Chatzinotas, S. Kisseleff, *et al.*, “Satellite communications in the new space era: A survey and future challenges,” *preprint arXiv:2002.08811*, 2020.

[151] Y. L., C. S., G. Y., H. H., W. J., Z. Y., and Y. S., “SatEC: A 5G Satellite Edge Computing Framework Based on Microservice Architecture,” *Sensors*, vol. 19, no. 4, p. 831, 2019.

[152] G. K., Q. M., Z. H., T. L., and Z. Z., “Dynamic energy-aware cloudlet-based mobile cloud computing model for green computing,” *J. Netw. Comput. Appl.*, 2016.

[153] L. L., C. Z., G. X., and M. S., “Multi-objective optimization for computation offloading in fog computing,” *IEEE Internet of Things J.*, 2018.

[154] W. N., V. B., M. M., and N. D., “Enorm: A framework for edge node resource management,” *IEEE Transactions on Services Computing*, 2017.

[155] T. Z., Y. F., L. X., J. H., and L. V., “Virtual resource allocation for heterogeneous services in full duplex-enabled SCNs with mobile edge computing and caching,” *IEEE Transactions on Vehicular Technology*, 2017.

[156] Y. C., H. K., C. H., and h. K., “Energy-efficient resource allocation for mobile-edge computation offloading,” *IEEE Transactions on Wireless Communications*, 2016.

[157] X. J. and R. S., “Online learning for offloading and autoscaling in renewable-powered mobile edge computing,” in *2017 Global Communications Conference*, 2017.

[158] L. J., S. W., and A. M., “Bandwidth-adaptive partitioning for distributed execution optimization of mobile applications,” *Academic Press Ltd*, 2014.

[159] W. J., H. J., and C. S., “Satellite IoT Edge Intelligent Computing: A Research on Architecture,” *Electronics*, vol. 8, no. 11, p. 1247, 2019.

[160] S. J. Olivieri, J. Aarestad, L. H. Pollard, A. M. Wyglinski, C. Kief, and R. S. Erwin, “Modular FPGA-based software defined radio for CubeSats,” in *2012 IEEE International Conference on Communications (ICC)*, pp. 3229–3233, IEEE, 2012.

[161] K. Varnavas, W. H. Sims, and J. Casas, “The use of field programmable gate arrays (FPGA) in small satellite communication systems,” 2015.

[162] J. Budroweit, “Design of a highly integrated and reliable SDR platform for multiple RF applications on spacecrafts,” in *GLOBECOM 2017-2017 IEEE Global Commu- nications Conference*, pp. 1–6, IEEE, 2017.

[163] J. Downey and T. Kacpura, “Pre-flight testing and performance of a Ka-band software defined radio,” in *30th AIAA International Communications Satellite System Conference (ICSSC)*, p. 15258, 2012.

[164] M. R. Maheshwarappa, *Software defined radio (SDR) architecture for concurrent multi-satellite communications*. University of Surrey (United Kingdom), 2016.

[165] “Enabling AI Research for 5G Networks with NI SDR,” white paper, National Instruments, 2020.

[166] “DARPA AI Colloquium,” Accessed: 2020-08-08.

[167] M. Azarmehr, A. Mehta, and R. Rashidzadeh, “Wireless device identification using oscillator control voltage as RF fingerprint,” in *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 1–4, IEEE, 2017.

[168] S. U. Rehman, K. W. Sowerby, S. Alam, and I. Ardekani, “Radio frequency fingerprinting and its challenges,” in *2014 IEEE Conference on Communications and Network Security*, pp. 496–497, IEEE, 2014.

[169] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, “Flexran: A flexible and programmable platform for software-defined radio access networks,” in *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*, pp. 427–441, 2016.

[170] DVB, “Second generation framing structure, channel coding and modulation systems for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications,” 2014. v1.4.1.

[171] DVB, “Second Generation DVB Interactive Satellite System (DVB-RCS2),” 2014. v1.2.1.

[172] S. Abdellatif, P. Berthou, P. Gelard, T. Plesse, and S. El-Yousfi, “Exposing an openflow switch abstraction of the satellite segment to virtual network operators,” in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, IEEE, 2016.

[173] 3GPP, “User Equipment (UE) radio transmission and reception,” TS 38.101, 3GPP, 2019. v15.5.0.

[174] P. S. Khodashenas, H. Khalili, D. Guija, and S. Siddiqui, “Talent: Towards integration of satellite and terrestrial networks,” in *2019 European Conference on Networks and Communications (EuCNC)*, pp. 167–171, IEEE, 2019.

[175] 3GPP, “Study on New Radio (NR) to support non-terrestrial networks,” TR 38.811, 3GPP, 2020. v15.3.0.

[176] 3GPP, “Solutions for NR to support non-terrestrial networks (NTN),” TR 38.821, 3GPP, 2019. v16.0.0.

[177] 3GPP, “Study on scenarios and requirements for next generation access technolo- gies,” TSG RAN TR38.913 R14, Jun. 2017.

[178] B. Soret, S. Ravikanti, and P. Popovski, “Latency and timeliness in multi-hop satellite networks,” in *Proc. IEEE ICC*, 2020.

[179] T. Li, H. Zhou, H. Luo, W. Quan, and S. Yu, “Modeling software defined satellite networks using queueing theory,” in *Proc. IEEE ICC*, 2017.

[180] K. Rusek, J. Suárez-Varela, P. Almasan, P. Barlet-Ros, and A. Cabellos-Aparicio, “RouteNet: Leveraging graph neural networks for network modeling and optimiza- tion in SDN,” *IEEE J. Sel. Areas Commun., early access*, 2020.

[181] B. Soret and D. Smith, “Autonomous routing for LEO satellite constellations with minimum use of inter-plane links,” in *Proc. IEEE ICC*, 2019.

[182] P. Almasan, J. Suárez-Varela, A. Badia-Sampera, K. Rusek, P. Barlet-Ros, and A. Cabellos-Aparicio, “Deep reinforcement learning meets graph neural networks: exploring a routing optimization use case,” *arXiv preprint arXiv:1910.07421*, 2020.

[183] M. Eisen and A. R. Ribeiro, “Optimal wireless resource allocation with random edge graph neural networks,” *IEEE Trans. Signal Process.*, vol. 68, pp. 2977–2991, 2020.

[184] Z. Gu, C. She, W. Hardjawana, *et al.*, “Knowledge-assisted deep reinforcement learning in 5G scheduler design: From theoretical framework to implementation,” *submitted to IEEE Jounral for possible publication*, 2020.

[185] X. Jiang, H. Shokri-Ghadikolaei, G. Fodor, E. Modiano, Z. Pang, M. Zorzi, and C. Fischione, “Low-latency networking: Where latency lurks and how to tame it,” *Proc. IEEE*, pp. 280–306, Feb. 2019.

[186] C. Sun, C. She, and C. Yang, “Unsupervised deep learning for optimizing wireless systems with instantaneous and statistic constraints,” *submitted to IEEE Jounral for possible publication*, 2020.

[187] R. Dong, C. She, W. Hardjawana, Y. Li, and B. Vucetic, “Deep learning for radio resource allocation with diverse quality-of-service requirements in 5g,” *IEEE Trans. Wireless Commun., minor revision*, 2020.

[188] E. Glaessgen and D. Stargel, “The digital twin paradigm for future NASA and US air force vehicles,” in *AIAA/ASME/ASCE/AHS/ASC Struct. Dyn. Mater. Conf.*, 2012.

[189] C. She and C. Yang, “Energy efficient resource allocation for hybrid services with future channel gains,” *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 1, pp. 165– 179, 2020.

[190] Z. Hou, C. She, Y. Li, Z. Li, and B. Vucetic, “Prediction and communication co-design for ultra-reliable and low-latency communications,” *IEEE Trans. Wirelss Commun.*, vol. 19, pp. 1196–1209, Feb. 2020.

[191] C. A. Balanis, *Antenna Theory: Analysis and Design*. USA: Wiley-Interscience, 2005.

[192] A. Alkhateeb, R. W. Heath, and G. Leus, “Achievable rates of multi-user millimeter wave systems with hybrid precoding,” in *2015 IEEE International Conference on Communication Workshop (ICCW)*, pp. 1232–1237, 2015.

[193] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, “A vector-perturbation technique for near-capacity multiantenna multiuser communication-part i: channel inversion and regularization,” *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 195–202, 2005.

[194] D. Nguyen and L. B. Le, “Resource allocation for multibeam miso satellite systems: Sum rate versus proportional fair optimization,” in *2016 IEEE Wireless Communications and Networking Conference*, pp. 1–6, IEEE, 2016.

[195] W. Wang, A. Liu, Q. Zhang, L. You, X. Gao, and G. Zheng, “Robust multigroup multicast transmission for frame-based multi-beam satellite systems,” *IEEE Access*, vol. 6, pp. 46074–46083, 2018.

[196] Y. Yan, W. Yang, B. Zhang, D. Guo, and G. Ding, “Outage constrained robust beamforming for sum rate maximization in multi-beam satellite systems,” *IEEE Communications Letters*, vol. 24, no. 1, pp. 164–168, 2019.

[197] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, “Transmit beamforming for physical-layer multicasting,” *IEEE transactions on signal processing*, vol. 54, no. 6, pp. 2239–2251, 2006.

[198] K. Shen and W. Yu, “Fractional programming for communication systems—part i: Power control and beamforming,” *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616–2630, 2018.

[199] S. Huaizhou, R. V. Prasad, E. Onur, and I. Niemegeers, “Fairness in wireless net- works: Issues, measures and challenges,” *IEEE Communications Surveys & Tuto- rials*, vol. 16, no. 1, pp. 5–24, 2013.

[200] G. Yuan and B. Ghanem, “An exact penalty method for binary optimization based on mpec formulation.,” in *AAAI*, pp. 2867–2875, 2017.

[201] I. Thibault, F. Lombardo, E. A. Candreva, A. Vanelli-Coralli, and G. E. Corazza, “Coarse beamforming techniques for multi-beam satellite networks,” in *2012 IEEE International Conference on Communications (ICC)*, pp. 3270–3274, IEEE, 2012.

[202] V. Joroughi, B. Devillers, M. Á . Vázquez, and A. Pérez-Neira, “Design of an on-board beam generation process for the forward link of a multi-beam broadband satellite system,” in *2013 IEEE Global Communications Conference (GLOBECOM)*, pp. 2921–2926, IEEE, 2013.

[203] V. Joroughi, M. B. Shankar, S. Maleki, S. Chatzinotas, J. Grotz, and B. Ottersten, “On-board precoding in a multiple gateway multibeam satellite system,” in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pp. 1–5, IEEE, 2018.

[204] A. Alkhateeb, G. Leus, and R. W. Heath, “Limited feedback hybrid precoding for multi-user millimeter wave systems,” *IEEE transactions on wireless communications*, vol. 14, no. 11, pp. 6481–6494, 2015.

[205] L. You, K.-X. Li, J. Wang, X. Gao, X.-G. Xia, and B. Ottersten, “Massive mimo transmission for leo satellite communications,” *arXiv preprint arXiv:2002.08148*, 2020.

[206] K.-X. Li, L. You, J. Wang, X. Gao, C. G. Tsinos, S. Chatzinotas, and B. Ottersten, “Downlink transmit design in massive mimo leo satellite communications,” *arXiv preprint arXiv:2008.05343*, 2020.

[207] M.-H. You, S.-P. Lee, and Y. Han, “Adaptive compensation method using the prediction algorithm for the doppler frequency shift in the leo mobile satellite communication system,” *ETRI journal*, vol. 22, no. 4, pp. 32–39, 2000.

[208] R. Hadani, S. Rakib, M. Tsatsanis, A. Monk, A. J. Goldsmith, A. F. Molisch, and R. Calderbank, “Orthogonal time frequency space modulation,” pp. 1–6, 2017.

[209] F. Hlawatsch and G. Matz, *Wireless communications over rapidly time-varying channels*. Academic press, 2011.

[210] P. Raviteja, K. T. Phan, Y. Hong, and E. Viterbo, “Interference cancellation and iterative detection for orthogonal time frequency space modulation,” vol. 17, pp. 6501–6515, Oct. 2018.

[211] W. Yuan, Z. Wei, J. Yuan, and D. W. K. Ng, “A simple variational Bayes detector for orthogonal time frequency space (OTFS) modulation,” vol. 69, no. 7, pp. 7976– 7980, 2020.

[212] A. A. Saleh, "Frequency-independent and frequency-dependent nonlinear models of twt amplifiers," *IEEE Transactions on communications*, vol. 29, no. 11, pp. 1715–1720, 1981.

[213] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of rf power amplifiers," *IEEE Transactions on signal processing*, vol. 54, no. 10, pp. 3852–3860, 2006.

[214] MathWorks, *5G Development with MATLAB*. MathWorks, 2020.

[215] MathWorks, *Deploying 5G NR Wireless Communications on FPGAs: A Complete MATLAB and Simulink Workflow*. MathWorks, 2020.

[216] MathWorks, *NR Cell Search and MIB and SIB1 Recovery*. MathWorks, 2020.

USER INFORMED **INDUSTRY** DRIVEN **RESEARCH** POWERED

SMARTSAT
COOPERATIVE RESEARCH CENTRE

**Australia's
Premier
Space
Research
Centre**

Australian Government
Department of Industry,
Science and Resources

AusIndustry
Cooperative Research
Centres Program

SmartSat CRC Head Office:
Lot Fourteen, Level 2, McEwin Building
North Terrace, Adelaide, SA

info@smartsatcrc.com
smartsatcrc.com