ITUJournal
Future and evolving technologies

ITU Journal on Future and Evolving Technologies, Volume 7, Issue 2, June 2026

ARTICLE

AI-driven orchestration for integrated satellite-terrestrial 6G networks: Architecture, handover management, and testbed evaluation

Murat Parlakisik [1,2]***, Ertan Ozturk*** [2]***, Nazlı Güney*** [3,4]

[1] *Amazon, Seattle, WA, USA,* [2] *University of North Dakota, Grand Forks, ND, USA,* [3] *TT Mobil Iletisim Hizmetleri, Istanbul, Turkiye,* [4] *Istanbul Rumeli University, Istanbul, Turkiye*

Corresponding author: Murat Parlakisik, murat.parlakisik@und.edu

The integration of terrestrial and Non-Terrestrial Networks (NTNs) is a cornerstone of the sixth-Generation (6G) vision, yet orchestrating artificial intelligence (AI) workloads across these heterogeneous domains remains an open challenge. This paper introduces an AI-driven orchestration architecture for integrated satellite-terrestrial 6G networks that extends four components of our previously proposed AI-native design, namely the 3rd Generation Partnership Project (3GPP) enhanced Network Data Analytics Function (NWDAF), Service Hosting Environment (SHE), Network Knowledge Exposure Function (NKEF), and AI agent framework, with NTN-specific capabilities. Drawing on a systematic analysis of the 20 ubiquitous connectivity use cases and the five AI-NTN convergence use cases from the 3GPP Technical Report (TR) 22.870 document, we conceptualize these architectural extensions: (i) an NTN-enhancement to NWDAF for ingesting satellite telemetry and ephemeris data to be used in predictive handover and coverage analytics, (ii) a space-edge computing tier within SHE that enables onboard satellite AI inference with a latency-aware model placement framework, (iii) a cross-domain AI agent for intent-driven orchestration across terrestrial and satellite operator boundaries, and (iv) an NKEF feature exposing constellation topology and coverage predictions to third-party applications. In order to demonstrate the effectiveness of this orchestration architecture, we present a handover management framework addressing four transition types (satellite-to-satellite, satellite-to-terrestrial hand-out, terrestrial-to-satellite hand-in, and inter-orbit) with backhaul-aware path selection leveraging inter-satellite links. Next, a comparative evaluation of eight NTN testbed platforms identifies current validation capabilities and gaps with respect to the discussed end-to-end system. Finally, we map the proposed architecture to 3GPP Releases 19–21, inspect five specification gaps that must be addressed to realize AI-driven TN-NTN orchestration, and describe a limitations and quantitative validation roadmap in the form of future work items.

Keywords: 6G architecture, AI-driven orchestration, handover management, non-terrestrial networks, NWDAF, satellite-terrestrial integration, space-edge computing, testbed evaluation

1. INTRODUCTION

The International Mobile Telecommunications-2030 (IMT-2030) framework, which is the sixth-Generation (6G) wireless system envisioned by the International Telecommunication Union (ITU), considers *ubiquitous connectivity* and *Artificial Intelligence (AI) and communication* as new usage scenarios, and positions *ubiquitous intelligence* as an overarching aspect [1]. Unlike fifth-Generation (5G) networks that treat AI as an overlay optimization, 6G is anticipated to embed AI capabilities directly into the network architecture, enabling autonomous operations, intelligent multi-agent ecosystems, and network-native AI processing [2, 3]. In support of the ubiquitous connectivity usage scenario of the IMT-2030, on the other hand, the proliferation of commercial Low Earth Orbit (LEO) constellations, which embrace direct-to-device or direct-to-cell technologies for unmodified cellular phones and Internet-of-Things (IoT) devices to connect directly to satellites that become cell towers in space, is accelerating the convergence of Terrestrial Networks (TNs) with Non-Terrestrial Networks (NTNs) to achieve a truly global coverage for communication service providers [4, 5].

Long before the official statement in IMT-2030 for a ubiquitous connectivity usage scenario, the 3rd Generation Partnership Project (3GPP) standardization body set its mission to progressively incorporate NTN support to 5G mobile networks that began with the foundational study in 3GPP Technical Report (TR) 38.811 [6], and culminated in the 3GPP Release 19 specifications frozen in December 2025 [7], which include NTN for New Radio (NR), IoT-NTN enhancements, inter-radio access technology (inter-RAT) mobility, and regenerative payload support. 3GPP Release 20, currently under development, introduces the "Satellite and NTN in 5G Architecture" work item [8], preparing the groundwork for normative 6G specifications expected in 3GPP Release 21 (2027–2028). In parallel, 3GPP TR 22.870 [9] defines sets of use cases for 6G such as the 61 AI use cases spanning network optimization, agent collaboration, and computing offloading, and 20 ubiquitous connectivity use cases explicitly involving NTN platforms, where five scenarios are identified across the complete list of use cases that require convergence of AI and NTN capabilities.

The NTN ecosystem covered in the 3GPP discussions encompasses three distinct platform classes operating at different altitudes and with fundamentally different mobility, coverage, and compute characteristics. Satellites (LEO at 500–2 000 km, MEO at 8 000 to 20 000 km, and GEO at 35 786 km) provide global or regional coverage but impose significant propagation delays and/or frequent handovers, particularly in LEO. High-Altitude Platform Stations (HAPS), operating quasi-stationarily in the stratosphere at approximately 20 km altitude, offer wide-area coverage with propagation delays as low as 0.06 ms and minimal Doppler shift, positioning them as intermediate nodes between satellites and terrestrial infrastructure [10, 11]. Unmanned Aerial Vehicles (UAVs), flying below 8 km, enable rapid on-demand deployment for localized coverage but face battery-limited endurance and three-dimensional mobility challenges that complicate handover management [12]. Recent surveys have examined the broader Space-Air-Gound Integrated Network (SAGIN) paradigm that unifies these three tiers [13]. While the full SAGIN vision requires orchestrating across all NTN platform types, this paper focuses specifically on the *satellite tier*, which presents the most demanding orchestration challenges due to high orbital velocities, long propagation delays, and the need for Inter-Satellite Link (ISL) routing through constellation meshes. The proposed architecture is designed to accommodate HAPS and UAV platforms in future extensions, and the use case analysis necessarily references all NTN platform types as defined by 3GPP, but the detailed architectural components, handover framework, and testbed validation target satellite-terrestrial integration.

With respect to implementation of standards-based integrated AI-native TN-satellite systems as targeted in this paper, orchestrating AI workloads across heterogeneous TN and satellite NTN domains poses fundamental challenges that neither pure terrestrial AI architectures nor conventional satellite management systems undertake separately. First, the compute continuum now extends from user devices through terrestrial edge and cloud to *space-edge* nodes on satellite payloads, requiring AI model placement decisions that account for power constraints, intermittent connectivity, and radiation effects. Second, LEO satellite visibility windows last only 5–10 minutes at typical altitudes and elevation angles, causing frequent inter-satellite handovers that demand predictive, AI-driven handover management that considers not only radio conditions but also ground-segment backhaul congestion and ISL routing alternatives. Third, cross-domain orchestration must coordinate between terrestrial Mobile Network Operators (MNOs) and satellite operators, each with distinct trust domains, service-level agreements, and network management systems.

Existing work addresses these challenges in isolation. Letaief et al. [2] and Saad et al. [3] provided early AI-for-6G roadmaps without NTN focus. Giordani and Zorzi [4] and Araniti et al. [14] surveyed NTN challenges for 6G but did not propose AI-native architectures. Recent studies have applied Deep Reinforcement Learning (DRL) to LEO handover optimization [15, 16] and explored space-edge computing [17], which lack unified AI architecture components such as the Network Data Analytics Function (NWDAF), Service Hosting Environment (SHE), and AI agents with NTN orchestration in a 3GPP-aligned framework validated on a realistic testbed. In our earlier work [18], we proposed a seven-component AI-native architecture derived from the 61 3GPP AI use cases, where we did not address NTN-specific extensions. In a series of other previous experimental work [19], we demonstrated an 86.7% Round-Trip Time (RTT) reduction through backhaul-aware handover in satellite-terrestrial integrated networks, which did not benefit from an AI orchestration logic.

This paper bridges these gaps by unifying and substantially extending our prior contributions into a single end-to-end framework. Relative to [18], which proposed the seven-component architecture for terrestrial 6G without NTN considerations, the present work introduces four NTN-specific architectural extensions with formal optimization formulations and timing analysis. In comparison to [19], which validated the testbed platform and backhaul-aware handover mechanism independently, this paper embeds those results within a unified NWDAF-centric orchestration framework and provides the architectural context for AI-native NTN management. The specific contributions are:

1. We extend the seven-component AI-native 6G architecture from [18] with NTN-specific modules: an NTN-enhanced NWDAF for satellite telemetry ingestion and predictive analytics, a space-edge computing tier

within SHE for onboard satellite AI, a cross-domain AI agent for multi-operator orchestration, and an NKEF extension for NTN knowledge exposure.
2. We present a systematic AI model placement framework for deploying models across the four-tier compute continuum (space-edge, ground-edge, regional edge, core cloud), analyzing the trade-offs among latency, power consumption, model complexity, and connectivity constraints.
3. We design an AI-driven handover management framework covering four TN-NTN transition types with backhaul-aware path selection that leverages NWDAF predictions and ISL routing.
4. We comparatively evaluate eight NTN testbed and simulation platforms, including our own solution, identifying their capabilities, limitations, and gaps for AI-native NTN orchestration validation.

The remainder of this paper is organized as follows. Section 2 presents the background on 3GPP AI use cases, NTN handover mechanisms, and AI model deployment challenges. Section 3 introduces the proposed AI-driven orchestration architecture. Section 4 details the handover management framework. Section 5 provides the testbed comparison and evaluation. Section 6 discusses standardization alignment and open challenges, as well as a quantitative validation roadmap for the architecture. Section 7 concludes with future research directions.

2. BACKGROUND AND RELATED WORK

2.1 3GPP use cases and NTN requirements

The forthcoming 6G standard foresees AI as a *native* design principle rather than an aftermarket optimization layer [1, 2]. To ground this vision in concrete service requirements, 3GPP TR 22.870 [9] catalogues an extensive set of use cases spanning enhanced mobile broadband, industrial automation, public safety, and aerial services. In our prior work [18], we systematically categorized 61 AI-related use cases from 3GPP TR 22.870 into ten functional domains, ranging from *network optimization* and *connected automated mobility* to *robotics and embodied AI* and *smart home and healthcare*, and proposed a seven-component AI-native architecture (SHE, NWDAF, NKEF, Integrated Sensing and Communication (ISAC), AI agent framework, federated learning, and semantic communication) to collectively satisfy the requirements of those domains. Four specification gaps were included as part of the study to point out the areas, for which 3GPP normative texts fall short of the envisioned AI-native operation. The present paper extends that analysis to the non-terrestrial dimension by examining the subset of 3GPP TR 22.870 use cases that explicitly involve satellite, HAPS, or UAV-based networks and, critically, the use cases where AI and NTN capabilities must converge.

3GPP TR 22.870 Section 8 defines 20 dedicated ubiquitous connectivity Use Cases (UCs). Among these, several impose demanding requirements on both the network architecture and AI functionality:

- UC 8.15 (Onboard Computing in 6G NTN Domain) focuses on the concept of space-edge computing, where AI model inference executes on satellite payloads. This reduces downlink bandwidth requirements for Earth-observation and weather-monitoring but introduces challenges related to limited power budgets, radiation-hardened computing, and thermal management.
- UC 8.17 (6G Satellite Backhaul) treats satellites, connected via high-capacity ISLs, as mesh backhaul infrastructure for remote terrestrial base stations, demanding Gbps-class throughput with Quality of Service (QoS) guarantees that rival fiber transport.
- UC 8.9 (Low-Altitude Logistics via NTN) is concerned with package-delivery drones operating beyond terrestrial coverage. NTN provides command-and-control, telemetry, and fleet management links whose reliability directly governs flight safety.
- UC 8.18 (Massive User Access in Disasters) addresses the scenario in which thousands of users contend for a single satellite beam during infrastructure-destroying events. Advanced access control, network slicing, and signaling-storm mitigation are required.

In addition, five cross-cutting use cases drawn from sections 6, 8, and 11 (AI, ubiquitous connectivity, and industry and verticals) blend AI and NTN capabilities. UC 6.14 (Intelligent UAV Swarms) defines a disaster-response scenario in which heterogeneous UAVs with varying compute and AI capabilities collaborate under network orchestration: The 6G system must discover AI models across UAV and edge nodes, perform distributed inference, and coordinate multi-UAV data collection within a 500 ms coordination latency budget. UC 11.5 (Immersive Media for Advanced Air Mobility) requires seamless terrestrial-to-satellite handover for electrical Vehicle Take-Off and Landing (eVTOL) aircraft passengers streaming 4K media during flight, with AI-driven coverage prediction and adaptive bit rate control. UC 11.7 (Assisted Airspace Management) relies on AI-based collision prediction with a 300 ms alert window, feeding a network digital twin of the airspace. UC 8.12 (Ubiquitous Emergency Rescue via UAVs) deploys UAVs over disaster areas using 6G NTN for command-and-control communication and leverages onboard computing in the operator-managed service hosting environment for intelligent rescue planning. UC 11.14 (Seamless Connectivity for 6G-Enabled Mission Critical Services) addresses first-responder scenarios where UEs transition between TN and NTN access (including satellite, UAV- and HAPS-based relay platforms), requiring mission critical service continuity across these heterogeneous access types.

Table 1 systematically categorizes all 20 ubiquitous con-

nectivity use cases from 3GPP TR 22.870 Section 8 into six functional groups, together with the five cross-cutting AI-NTN convergence use cases identified above. Note that UC 8.12 appears in both the Disaster & Emergency category (as one of the 20 Section 8 use cases) and the AI-NTN Convergence category (as one of the five cross-cutting use cases), yielding 24 distinct use cases across the table. The categorization reveals three structural observations. First, NTN positioning and timing constitutes the largest single group (six use cases), underscoring the importance of location services in the NTN domain. Second, latency requirements span two orders of magnitude across categories, from sub-100 ms for mission-critical services to 1000 ms for in-flight media streaming, implying that a one-size-fits-all AI model placement strategy is insufficient; orchestration must be *latency-aware*. Third, the majority of safety-critical use cases demand ≥99.9% reliability, which necessitates redundant TN-NTN connectivity paths and AI-driven failover mechanisms.

To position the proposed architecture against the relevant broader literature, we additionally consider three complementary research threads. First, NWDAF-based mobility analytics introduced in 3GPP Release 17 [20] provide standardized analytics IDs for UE mobility prediction, communication-pattern prediction, and QoS sustainability [21, 22]; however, these analytics are defined exclusively over terrestrial network functions, and the NTN-enhanced NWDAF proposed in Section 3.2 extends this baseline with satellite-specific data sources (ephemeris, ISL topology, ground-station load) that are not present in the Release 17/18 specification. Second, the Open-Radio Access Network (O-RAN) xApp/rApp paradigm for closed-loop RAN control has recently been extended to satellite-segment intelligence by the Space-O-RAN initiative [23] and to AI/ML-based anomaly detection through self-adaptive xApps [24]; however, in contrast to RAN Intelligent Controller (RIC)-centric approaches that operate at the RAN layer, the present work targets the core-network analytics function and the cross-domain agent layer, which sit one level above the RIC and consume RIC-produced telemetry as input (Section 2.3.3) Third, recent measurement studies of operational LEO constellations [25] confirm the latency profile assumed in our design (median ground-to-LEO RTT in the 40–60 ms range without ground-segment congestion, increasing markedly under load), and emulation platforms [26] calibrate their induced delays against the same empirical Starlink dataset, providing an empirical baseline against which the architectural assumptions in this paper can be validated.

2.2 NTN handover state of the art

Handover management in integrated satellite-terrestrial networks poses challenges that are qualitatively different from purely terrestrial mobility. Three scenario classes dominate the literature: *satellite-to-satellite* (intra-NTN), *satellite-to-terrestrial* (vertical), and *terrestrial-to-satellite* (reverse vertical). A fourth class, *inter-orbit* handover between LEO, MEO, and GEO layers, which has received comparatively little attention, is addressed as part of our handover framework in Section 4.1. Each class introduces distinct timing, signaling, and prediction requirements that we review in turn.

2.2.1 3GPP handover mechanisms for NTN

3GPP Release 17 introduced baseline NTN mobility support by extending the New Radio (NR) handover framework with satellite-specific enhancements [6]. *Conditional Handover* (CHO) allows the UE to be preconfigured with multiple candidate target cells and associated execution conditions, and hence when a condition is met, typically a measurement threshold on Reference Signal Received Power (RSRP) or Signal-to-Interference-plus-Noise Ratio (SINR), the UE autonomously executes the handover without further network signaling. For LEO constellations, CHO is augmented with *ephemeris-assisted* cell selection: The network provides satellite orbital parameters so that the UE can predict beam coverage windows and pre-compute target cell rankings. *Dual Active Protocol Stack* (DAPS) handover further reduces interruption time by maintaining simultaneous connections to source and target cells during the transition, at the cost of increased UE complexity and power consumption.

Despite these advances, two fundamental challenges remain. First, LEO satellite passes at altitudes of 500–2 000 km produce visible windows as short as 5–10 minutes, with maximum radial Doppler shifts of up to ±270 kHz at 28 GHz Ka-band for a 550 km orbit at typical operational elevation angles [6]. The resulting handover frequency, once every few minutes per UE in a dense constellation, places heavy demands on the core network signaling plane. Second, the propagation delay asymmetry between terrestrial links (<10 ms one-way) and LEO satellite links (25–50 ms one-way) introduces timing ambiguities that can cause premature or late handover execution when conventional measurement-report-based triggers are used.

2.2.2 AI/ML approaches to NTN handover

Recent literature has increasingly turned to Machine Learning (ML) to address the shortcomings of threshold-based handover. Liu et al. [15] proposed a DRL agent that selects handover timing and target satellite jointly, achieving a 64% reduction in handover frequency compared to A3-event-based triggering in a Walker-Delta LEO constellation. The agent observes RSRP, Doppler

Table 1 – Categorization of NTN use cases from 3GPP TR 22.870 [9]. Latency ranges are aggregated from per-use case potential requirements in [9]; the "Architecture Relevance" column is an interpretive mapping to the architectural components performed by the authors, not a normative claim.

Category	Count	Use Cases	Latency Range	Architecture Relevance
Section 8: Ubiquitous Connectivity Use Cases (20 total)				
NTN Positioning & Timing	6	8.5 Resilient positioning, 8.7 Low-energy positioning, 8.10 Hybrid TN-NTN positioning, 8.11 Hybrid NTN-GNSS positioning, 8.14 Time distribution, 8.16 Positioning integrity	10–100 ms	NWDAF ephemeris prediction, NKEF exposure
Coverage, Continuity & Coordination	6	8.2 Ubiquitous network, 8.3 Sparse LEO UX, 8.4 Wearable continuity, 8.8 Global mobile video, 8.20 Cooperating UEs, 8.21 Inter-PLMN coordination	50–1000 ms	Cross-domain AI Agent, NWDAF handover analytics
Disaster & Emergency Response	3	8.6 Disaster relief, 8.12 Emergency rescue via UAVs, 8.18 Massive satellite access in disasters	<500 ms	AI Agent orchestration, SHE space-edge computing
HAPS-Based Services	2	8.13 HAPS for public safety, 8.19 HAPS traffic offloading	<500 ms	NKEF HAPS exposure, aerial-edge placement
NTN Infrastructure & Computing	3	8.9 Low-altitude logistics via NTN, 8.15 Onboard NTN computing, 8.17 Satellite backhaul	<200 ms	SHE space-edge tier, ISL-aware routing
Cross-Cutting AI-NTN Convergence Use Cases (Sections 6, 8, 11)				
AI-NTN Convergence	5	6.14 Intelligent UAV swarms, 8.12 Emergency rescue via UAVs, 11.5 Immersive media for AAM, 11.7 Assisted airspace management, 11.14 Mission critical services	100–1000 ms	All four components: NWDAF, SHE, AI Agent, NKEF

rate, and remaining visible time, and learns a policy that implicitly trades ping-pong avoidance against signal degradation risk. Chen et al. [27] extended this paradigm by representing the dynamic satellite-terrestrial topology as a graph and training a Graph Neural Network (GNN) coupled with a DRL agent. By representing satellites and terrestrial cells as graph nodes connected by ISL and adjacency edges, the GNN captures the constellation topology and allows the DRL agent to evaluate *multi-hop* ISL rerouting paths during handover decisions, an ability that conventional flat state vectors cannot provide because a fixed-length list of per-link metrics (e.g., RSRP and Doppler per candidate satellite) encodes only direct access-link quality and carries no information about how satellites interconnect through ISLs or which ground stations are reachable via intermediate hops. This approach is particularly relevant for aeronautical 6G scenarios in which both the UE (aircraft) and the serving node (satellite) are mobile. Wang et al. [28] focused on the traffic dimension, using a Long Short-Term Memory (LSTM) network to predict per-beam traffic load 30 s ahead and feeding these predictions into an attention-enhanced Rainbow Deep Q-Network (DQN) for joint handover-and-load-balancing decisions. Their results show a 23% throughput improvement in congested beams relative to load-unaware DRL baselines.

Complementing these AI-centric approaches, a low-complexity hybrid strategy that combines ephemeris-based pre-filtering of candidate satellites with a lightweight neural classifier for final target selection is proposed by Vasquez et al. [16], demonstrating that even modest ML models can outperform purely geometric predictors when channel fading statistics are incorporated.

2.2.3 Backhaul-aware handover

A dimension largely overlooked in existing work is the *backhaul path* that connects the serving satellite (or terrestrial cell) to the core network. In LEO constellations with ISLs, the end-to-end latency experienced by the UE depends not only on the access link quality but also on the number of ISL hops and the congestion state of the satellite backhaul mesh. In our prior experimental work [19], we demonstrated that when the ground-station path traverses congested ISL hops, end-to-end RTT increases by up to 4× relative to a direct ground-station link. Incorporating ISL-hop count and per-link utilization into the handover decision yielded significant RTT improvements at the cost of only 12% additional signaling overhead.

2.2.4 Identified gap

Surveying the literature, we observe that existing AI-driven handover solutions operate in isolation from the broader 3GPP network data analytics framework. None of the reviewed work integrates the handover decision agent with NWDAF [20] or exposes handover-relevant analytics (beam load predictions, ISL congestion maps) through standardized service interfaces. Furthermore, no prior work combines a 3GPP-aligned AI-native architecture with backhaul-aware, ISL-topology-cognizant handover optimization. This paper addresses precisely this gap by embedding the handover decision engine within the NTN-enhanced NWDAF and cross-domain AI agent framework described in Section 3, and validated through testbed experiments (Section 5).

2.3 AI model deployment in NTN

Deploying AI models in non-terrestrial environments introduces a set of constraints that are absent in terrestrial edge computing. We organize the discussion along three deployment paradigms: *on-ground processing*, *onboard satellite computing*, and *hybrid space-ground* approaches.

2.3.1 On-ground processing

The conventional approach routes all satellite-collected data to ground stations for centralized AI inference. This paradigm benefits from unconstrained compute and power budgets and allows operators to leverage commercial Graphical Processing Unit (GPU) clusters for training and inference. However, it suffers from two fundamental bottlenecks. First, the downlink bandwidth of an LEO satellite is finite, typically 1–10 Gbps per ground-station pass, and the contact window is limited to 5–10 minutes per satellite, creating a data-backlog problem for bandwidth-intensive workloads such as Synthetic Aperture Radar (SAR) imagery. Second, the round-trip latency through the satellite-ground-satellite path (50–600 ms depending on orbit) precludes real-time decision loops required by use cases such as low-altitude UAV supervision involving collision prediction (UC 7.5 in TR 22.870 Section 7, <300 ms) and multi-sensor fusion for UAV take-off and landing (UC 7.16, 100–500 ms).

2.3.2 Onboard satellite computing

UC 8.15 of 3GPP TR 22.870 [9] explicitly envisions *space-edge computing*, in which satellite payloads host general-purpose processors capable of running AI inference workloads. Recent advances in radiation-tolerant System-on-Chip (SoC) platforms, such as Xilinx Versal and Intel Agilex deployed on European Space Agency's (ESA's) PhiSat missions, have demonstrated that convolutional neural network inference for cloud detection and image classification can execute onboard with power envelopes of 5–15 W [29]. Despite these advances, several challenges persist: (i) limited thermal dissipation in the vacuum environment constrains sustained compute throughput; (ii) single-event upsets caused by cosmic radiation require error-correcting memory and checkpoint-restart mechanisms for long-running inference; and (iii) model updates are constrained by the narrow uplink windows and the need for over-the-air integrity verification.

2.3.3 Hybrid space-ground approaches

The emerging consensus in the literature is that neither purely on-ground nor purely onboard processing is optimal; instead, a *tiered* deployment that splits the AI pipeline across space-edge, aerial-edge, and ground-edge nodes is required [17, 30]. Javaid et al. [17] explored the use of Large Language Models (LLMs) in integrated satellite-aerial-terrestrial networks, proposing a split-inference architecture in which lightweight encoder layers execute onboard the satellite while decoder layers run at ground stations. Chen et al. [30] surveyed AI-driven techniques for TN-NTN integration in 6G, highlighting the potential of Federated Learning (FL) across satellite-ground boundaries to keep raw data localized while aggregating model updates at the ground station. However, FL over NTN faces unique obstacles: (i) the intermittent and asymmetric connectivity between satellites and ground stations disrupts the synchronous aggregation rounds assumed by federated averaging algorithms [31]; (ii) data distributions across satellites in different orbital planes are inherently non-independent and identically distributed (non-i.i.d.), amplifying convergence issues that techniques such as federated proximal (FedProx) algorithms [32] only partially mitigate; and (iii) Byzantine robustness is critical because a compromised satellite node could poison the global model without physical inspection being feasible.

The Space-O-RAN initiative [23] has proposed an O-RAN-aligned architecture for NTN in which RICs are instantiated at both ground and satellite segments, enabling AI model lifecycle management, including deployment, monitoring, and replacement, through standardized interfaces. While promising, this work focuses on the RAN layer and does not address core-network analytics (NWDAF) or cross-domain orchestration spanning multiple operator domains. Our proposed architecture differs from Space-O-RAN in three key aspects: (i) we extend the 3GPP core-network analytics function (NWDAF) with NTN-specific data sources and prediction capabilities, rather than operating solely at the RAN layer; (ii) we introduce a cross-domain AI agent that coordinates across operator trust boundaries, whereas Space-O-RAN as-

sumes a single-operator deployment; and (iii) we formalize a four-tier compute continuum (space-edge through core cloud) with an optimization-based model placement framework, whereas Space-O-RAN focuses on RIC placement without addressing AI workload distribution. The ETSI NTN vision paper [33] similarly outlines NTN integration requirements for 6G but remains at a high-level requirements stage without proposing concrete architectural components or orchestration mechanisms.

In this paper, we extend the Service Hosting Environment (SHE) component of our architecture [18] with a dedicated *space-edge tier* that formalizes the compute, storage, and AI model hosting capabilities available on regenerative satellite payloads, and we define the interfaces through which the NTN-enhanced NWDAF ingests satellite telemetry and ephemeris data for predictive analytics (Section 3.2).

3. AI-DRIVEN ORCHESTRATION ARCHITECTURE FOR TN-NTN INTEGRATION

3.1 Architecture overview

The proposed architecture extends our seven-component AI-native 6G framework [18] to address the unique requirements of integrated satellite-terrestrial operation. The original architecture, comprising SHE, NWDAF, NKEF, ISAC, AI agent framework, federated learning, and semantic communication, was derived from cross-category analysis of 61 3GPP AI use cases but assumed a predominantly terrestrial deployment. The extensions introduced here address three structural deficiencies that arise when NTN platforms enter the orchestration scope.

First, the compute continuum must expand beyond device-edge-cloud to include a *space-edge* tier on satellite payloads, fundamentally altering AI model placement decisions. Second, the analytics function (NWDAF) must ingest satellite-specific telemetry, including orbital position, link budget, ISL topology, and ground station load, and perform *ephemeris-based prediction* that exploits the deterministic nature of satellite orbits. Third, orchestration must operate across trust boundaries between terrestrial MNOs and satellite operators, requiring cross-domain AI agents with appropriate authentication and coordination protocols.

Fig. 1 illustrates the proposed architecture organized into three orchestration domains. The *Terrestrial Domain* contains conventional 6G infrastructure: AI-enabled advanced Node Bs (aNBs), terrestrial edge servers, and the core network with 5G/6G network functions. The *Non-Terrestrial Domain* encompasses LEO/MEO/GEO satellites with optional onboard computing (space-edge), HAPS, ground stations with co-located edge servers, and ISL mesh connectivity. The *Cross-Domain Orchestration Layer* bridges both domains through four AI-native components: NTN-enhanced NWDAF (Section 3.2), Extended SHE with space-edge tier (Section 3.3), Cross-domain AI agent (Section 3.4), and NKEF for NTN (Section 3.5). The orchestration layer operates on five design principles:

1. *Cross-domain awareness*: All orchestration decisions jointly consider TN and NTN topology, propagation characteristics, and available capacity.
2. *Predictive intelligence*: Ephemeris-based satellite trajectory prediction enables proactive handover preparation 30–60 seconds before coverage changes occur.
3. *Backhaul-aware management*: Ground-segment congestion is treated as a first-class optimization variable alongside radio conditions, enabling ISL-assisted rerouting [19].
4. *Zero-touch operation*: Closed-loop automation with (predict→decide→execute→monitor) eliminates manual intervention for routine TN-NTN transitions [34].
5. *Graceful degradation*: When TN-NTN transitions cause transient QoS degradation, the architecture adapts service parameters (bit rate, resolution, update frequency) rather than dropping sessions.

3.2 NTN-enhanced NWDAF

The 3GPP NWDAF, standardized in TS 23.288 [20], provides AI-native network analytics with data collection, model training and inference, and exposure interfaces. In its current form, NWDAF collects telemetry from RAN and core network functions and produces analytics consumed by other Network Functions (NFs) or applications. For TN-NTN integration, we extend NWDAF with three NTN-specific capabilities.

The NTN-enhanced NWDAF introduces a new satellite telemetry ingestion interface that collects: (i) orbital position and velocity vectors from satellite tracking systems or Two-Line Element (TLE) sets, (ii) per-satellite link budget parameters (transmit power, antenna gain, atmospheric attenuation), (iii) ISL topology and per-link utilization, and (iv) ground station backhaul capacity and congestion metrics. This telemetry is collected via standardized NWDAF data collection interfaces or monitoring exporters extended for NTN data sources [35]. The ingestion rate must accommodate LEO dynamics: At 550 km altitude with ~7.5 km/s orbital velocity, satellite geometry changes materially within seconds, requiring sub-second telemetry updates for safety-critical applications and 5–10 second intervals for routine optimization.

Unlike terrestrial UE mobility, which is stochastic, satellite trajectories are deterministic and predictable from orbital mechanics. The NTN-enhanced NWDAF exploits this by implementing Simplified General Perturbations-4

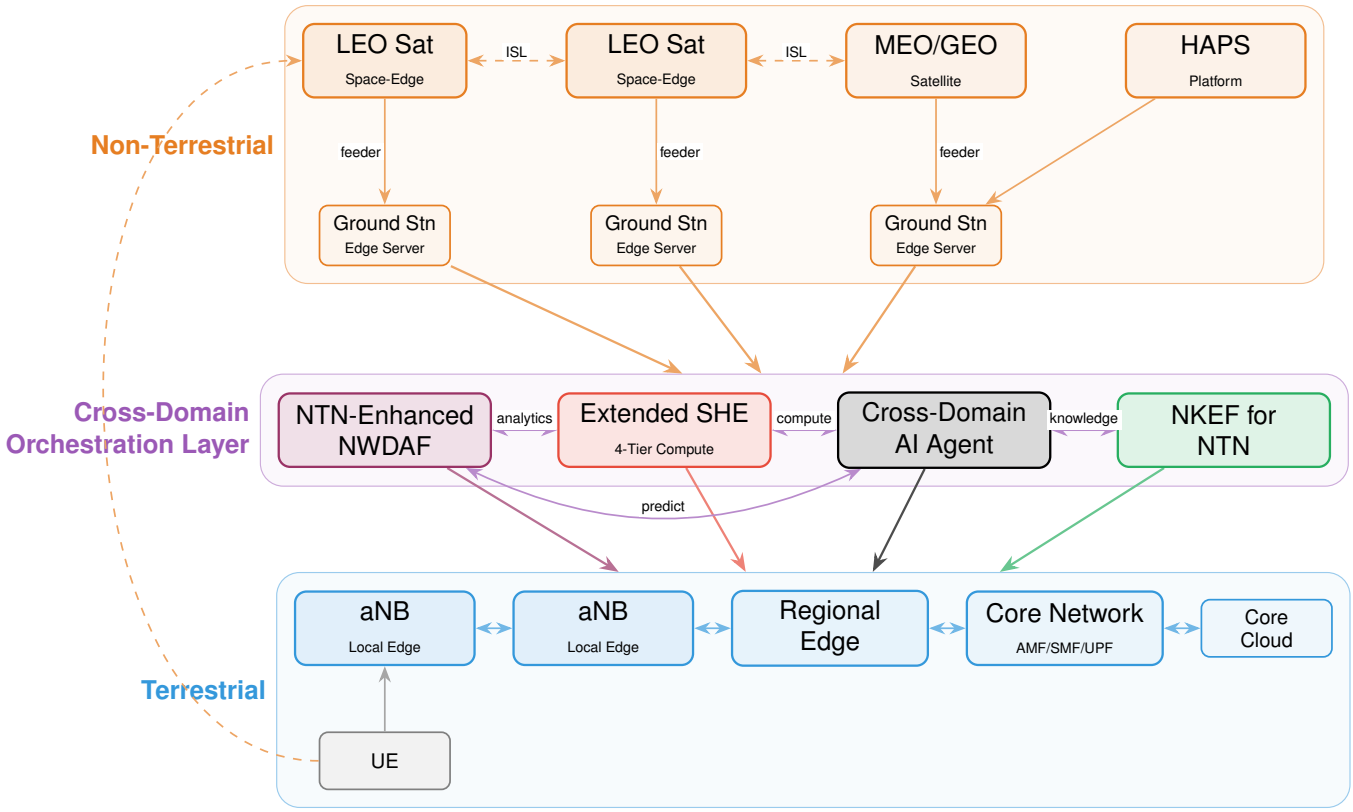

Figure 1 – AI-driven orchestration architecture for integrated satellite-terrestrial 6G networks.

(SGP4) propagation to predict satellite positions 30–60 seconds ahead. This enables: (a) coverage window prediction, which determines when a satellite will enter or exit a UE's field of view; (b) handover timing prediction, which computes the optimal handover trigger point that minimizes service interruption; (c) ground station visibility forecasting, which identifies which ground stations will be in line-of-sight of a given satellite over the next pass; and (d) ISL path pre-computation, which determines optimal ISL routing before handover execution.

The enhanced NWDAF operates in a closed-loop automation cycle: it *predicts* upcoming satellite geometry changes and ground station load trends, the AI agent *decides* on the optimal handover action (Section 4.2), the orchestration system *executes* the decision through the handover module, and telemetry metrics *monitor* the post-handover performance. The observed outcomes feed back into the analytics models for continuous refinement. This cycle operates at two timescales: a fast loop (seconds) for real-time handover decisions and a slow loop (minutes to hours) for model retraining and policy updates.

The accuracy of SGP4-based prediction is not assumed to be perfect, and the proposed architecture explicitly accommodates ephemeris uncertainty in two ways. First, SGP4 propagation accuracy degrades as the prediction horizon lengthens: published TLE-based predictions exhibit along-track errors growing from a few hundred meters for sub-hour horizons to several kilometers over 24–72 hours due to unmodelled atmospheric drag, J_2 and higher-order geopotential terms, and solar radiation pressure [29]. Because the handover preparation window in our design is bounded by the visibility horizon (typically 1–3 s, with handover decisions taken 30–60 s ahead of the geometric pass boundary as discussed in Section 4.2), the operating regime is dominated by sub-hour propagation errors, which map to timing errors below 100 ms. A 100 ms timing error translates into an angular position uncertainty below 1 km at LEO velocities, well within the beam-footprint margin needed to preserve handover correctness; however, when ephemeris age exceeds 12–24 hours, the cumulative drift can shift the predicted entry/exit moment of a pass by several seconds, increasing the probability of premature triggering (handover-too-early, leading to performing a Random Access Channel (RACH) procedure on a target whose beam has not yet illuminated the UE) or late triggering (handover-too-late, leading to radio-link failure on the source). Empirical evaluations of the graph-based handover paradigm originally introduced by Hu et al. [36] have reported reductions in handover-failure rate from above 5% to below 1% when ephemeris updates are refreshed at intervals consistent with our sub-second-to-10-second ingestion rate [37]. Second, to mitigate the residual prediction error, the NWDAF closed-loop monitor uses post-handover RACH-success and radio-link-failure counts to detect

drift and triggers re-propagation from a fresh TLE or, if available, on-board GNSS-derived ephemeris. Atmospheric scintillation, ionospheric delay variation at Ka-band, and rain-fade events are not deterministic, and are handled through the stochastic shadowing model in the link-budget block of the NWDAF rather than through the orbital-mechanics block. A complete quantitative characterization of the timing-error-to-handover-failure mapping under realistic perturbation distributions is left to future experimental work on the testbed described in Section 5.

3.3 SHE extension: Space-edge Tier

In our prior work [18], SHE was defined as a three-tier distributed computing infrastructure for terrestrial 6G: *local edge* (co-located with base stations, <10 ms latency), *regional edge* (aggregation points, <20 ms), and *core cloud* (centralized data centers). For TN-NTN integration, we extend this to a four-tier compute continuum by introducing a *space-edge* tier for onboard satellite computing and generalizing the original local edge into a *ground-edge* tier that encompasses both terrestrial base stations and satellite ground stations, as motivated by UC 8.15 of 3GPP TR 22.870 [9].

The extended SHE enables AI model placement across a four-tier compute continuum, each tier with distinct characteristics:

- The *space-edge* tier runs on onboard satellite processors and is suitable for lightweight models such as anomaly detection, basic signal classification, and link quality estimation. It is constrained by limited power budgets (solar panels and battery), radiation-hardened processors with reduced compute capacity compared to terrestrial GPUs, thermal management in vacuum, and intermittent ground connectivity (visibility windows of 5–10 minutes per ground station pass for LEO). Current space-edge compute capacity ranges from tens of Giga Operations per Second (GOPS) on radiation-tolerant processors to up to 275 Tera OPS (TOPS) on commercial modules in radiation-shielded enclosures [38, 39], with next-generation space-qualified AI accelerators projected to reach hundreds of TOPS by the 6G deployment timeframe.
- The *ground-edge* tier generalizes the original local edge from [18] to encompass both terrestrial base stations (aNBs) and satellite ground stations. At terrestrial base stations, this tier serves latency-sensitive AI workloads (<10 ms) such as beam management, channel estimation, and local inference for connected vehicles and IoT devices. At satellite ground stations, it hosts medium-complexity models including handover decision agents and real-time video analytics, benefiting from low latency to satellites in view (<10 ms for LEO at 550 km) and unconstrained power. Compute capacity at this tier typically ranges from hundreds of TOPS to hundreds of trillion floating-point operations per second (TFLOPS) per node, with power budgets of 150–300 W. In integrated TN-NTN scenarios, the ground-edge also acts as the termination point for satellite-to-terrestrial handovers, receiving migrated application state from the space-edge tier when a UE transitions to terrestrial coverage.
- The *regional edge* at terrestrial aggregation points supports complex models such as LLM inference, multi-sensor fusion, federated learning aggregation, and video analytics pipelines. It connects to ground stations via terrestrial backhaul. Typical deployments provide aggregate compute capacity of tens of Peta FLOPS per site through multi-GPU server clusters, with power envelopes of 5–20 kW per rack.
- The *core cloud* in centralized data centers is reserved for model training, long-term analytics, non-real-time generative AI, and global constellation management. It has the highest latency but virtually unlimited compute capacity, with hyperscale deployments capable of training billion-parameter models at power budgets of hundreds of kilowatts to megawatts per facility.

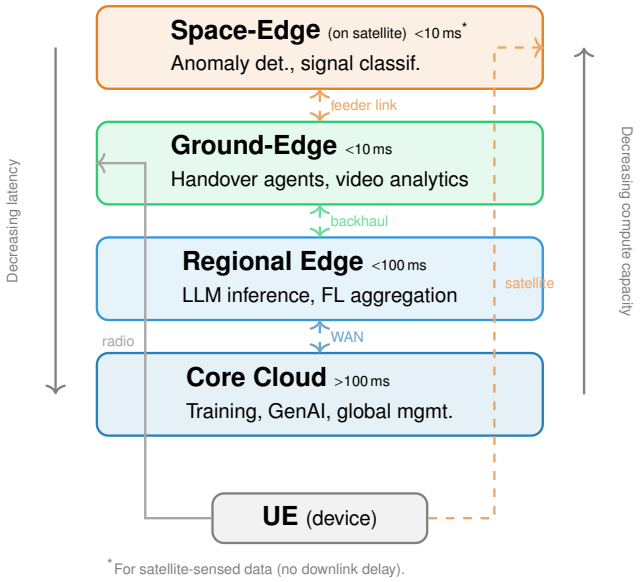

Figure 2 – Four-tier SHE compute continuum for AI model placement.

Fig. 2 illustrates the four-tier compute continuum, showing the trade-off between latency and compute capacity across tiers: space-edge offers the lowest latency for satellite-sensed data but the most constrained compute, while core cloud provides unlimited capacity at the cost of the highest latency. The orchestration layer selects the optimal tier for each AI workload based on five criteria: (i) *latency requirement*, where safety-critical tasks (UC 7.16 sensor fusion: <100 ms) demand ground-edge or space-edge while non-real-time analytics tolerate core cloud; (ii) *model complexity*, where parameter count and FLOPS determine minimum compute capability; (iii) *data locality*,

Algorithm 1 AI model deployment across four-tier SHE for integrated TN-NTN

Require: Model m with compute requirement C_m (FLOPS), energy cost E_m (W), latency requirement L_m, data source $d_m \in$ {satellite-sensed, terrestrial}
Require: Per-tier state: compute capacity C_t, available energy E_t, connectivity status $\sigma_t \in \{0, 1\}$ for each tier t
Ensure: Deployment tier $t^* \in$ {space-edge, ground-edge (GS/BS), regional edge, core cloud}
▷ Model weights are pre-positioned from core-cloud to all candidate tiers
1: **if** d_m = satellite-sensed **then** ▷ NTN data path
2: **if** $L_m < 10$ ms **and** $C_m \le C_{\text{space-edge}}$ **and** $E_m \le E_{\text{space-edge}}$ **then**
3: $t^* \leftarrow$ space-edge ▷ onboard satellite inference
4: **else**
5: $t^* \leftarrow$ ground-edge (GS) ▷ downlink to ground station
6: **end if**
7: **else if** d_m = terrestrial **then** ▷ TN data path
8: **if** $L_m < 10$ ms **then**
9: $t^* \leftarrow$ ground-edge (BS) ▷ at base station
10: **else if** $L_m < 100$ ms **then**
11: $t^* \leftarrow$ regional-edge ▷ aggregation point
12: **else**
13: $t^* \leftarrow$ core-cloud ▷ centralized DC
14: **end if**
15: **end if**
▷ Inter-orbit: if serving orbit changes, migrate model state
16: **if** inter-orbit handover triggered (LEO ↔ MEO/GEO) **then**
17: Transfer model state from current space-edge to new orbit's space-edge or ground-edge during visibility window
18: **end if**
▷ Fallback: selected tier unreachable or resource-insufficient
19: **if** $\sigma_{t^*} = 0$ **then**
20: $\mathcal{F} \leftarrow \{t \mid \sigma_t = 1, C_t \ge C_m, E_t \ge E_m\}$
21: **if** $\mathcal{F} \neq \emptyset$ **then**
22: $t^* \leftarrow \arg\min_{t \in \mathcal{F}} \text{latency}(t)$
23: **else**
24: $t^* \leftarrow$ degrade to lighter model m' with $C_{m'} < C_m$; retry
25: **end if**
26: **end if**
27: **return** t^*

where satellite-sensed data processed onboard avoids costly downlink bandwidth; (iv) *connectivity availability*, where space-edge processing becomes the only option when no ground station is in direct view or when the available ISL path to a reachable ground station exceeds the latency budget required by the service-level agreement; and (v) *energy budget*, where onboard processing is constrained by the satellite's power generation and storage capacity. Algorithm 1 formalizes this decision logic using latency thresholds derived from the use case requirements in Table 1. In production deployments, these thresholds would be configurable per service class and could be refined through the NWDAF's closed-loop monitoring.

The deployment proceeds as follows. The orchestrator first classifies the workload by data source. For satellite-sensed data (e.g., Earth-observation imagery, onboard telemetry), it attempts space-edge placement if the model fits within the satellite's compute capacity ($C_m \le C_{\text{space-edge}}$) and power budget ($E_m \le E_{\text{space-edge}}$), and the latency requirement is under 10 ms; otherwise, the data is downlinked to the co-located ground station for inference. For terrestrial data (e.g., UE measurements, IoT sensor feeds), the algorithm routes the workload through the conventional edge hierarchy: base-station ground-edge for sub-10 ms tasks, regional edge for sub-100 ms, and core cloud for the rest. A connectivity check (σ_{t^*}) then verifies that the selected tier is reachable; if not, for instance because no ground station is in view and the ISL path exceeds the SLA latency budget, the algorithm falls back to the lowest-latency reachable tier that satisfies both the compute and energy constraints. Model weights are pre-positioned from the core cloud to candidate tiers, which takes place during ground station visibility and/or ISL availability windows for the space-edge tier, so that inference can begin immediately upon workload arrival.

As satellites traverse their orbits, the optimal processing tier may change. For instance, an Earth-observation model running on an LEO satellite's space-edge processor must hand off its state to a ground-edge server when the satellite approaches the ground station horizon, allowing continued processing with higher compute capacity. Conversely, when a satellite can no longer reach any ground station, either directly or via ISL paths within the required SLA latency budget, pre-trained lightweight models must be available onboard for autonomous operation. The SHE orchestrator manages these migrations through three mechanisms: (i) saving the model's current state (weights and intermediate results) on the source tier and transferring it to the target tier via the satellite downlink during visibility windows, so that inference can resume without restarting from scratch; (ii) pre-positioning model weights on satellites expected to serve upcoming coverage areas before they lose ground connectivity; and (iii) graceful degradation to simpler, lower-complexity models when the full model state cannot be transferred within the available visibility window.

Algorithm 1 performs a static feasibility check and abstracts two dynamic costs. (i) *Migration cost*: Transferring a model of state size S_m across n_{hops} ISL paths, each with delay τ_{ISL} and bandwidth BW_{ISL}, requires $T_{\text{mig}} \approx S_m/BW_{\text{ISL}} + n_{\text{hops}} \cdot \tau_{\text{ISL}}$, which must satisfy $T_{\text{mig}} < T_{\text{vis}}$ (the migration time < the remaining visibility window); the feeder-link capacity imposes the same constraint on cold-start weight pre-positioning, so model uploads are scheduled at the start of a ground-station pass. (ii) *Resource contention* arises when $\sum_m C_m > C_{\text{space-edge}}$ or $\sum_m E_m > E_{\text{space-edge}}$; arbitration follows 5G QoS Identifier (5QI) service-class priority (latency-sensitive workloads pre-empt best-effort) and onboard power budget (housekeeping and station-keeping pre-empt AI inference, with throttling or migration to ground-edge during eclipse or manoeuvres). Both costs surface in the algorithm's fallback branch (line 18) and the graceful-degradation step. A full constrained queueing or knapsack formulation, with empirical contention and migration-completion characterisation, appears as Roadmap item R2 in Section 6.3.

Several open challenges remain for space-edge deployment [40, 41]. Radiation-induced single-event upsets can corrupt model weights and inference results, requiring error-correcting codes or triple modular redundancy. Model updates must be distributed across constellations of thousands of satellites, creating a massive model logistics problem. Power-aware inference scheduling must balance AI workload execution against communication duties and housekeeping operations. Finally, the lack of standardized APIs for space-edge compute resource discovery and allocation (identified as Gap 2 in Section 6.2) hinders interoperability.

3.4 Cross-domain AI agent framework for NTN orchestration

AI agents, defined as autonomous entities with identities, capabilities, goals, and LLM-based reasoning [18, 42, 43], form the largest category of 3GPP AI use cases (24.6%, 15 use cases). In the TN-NTN context, the AI agent framework must be extended to support cross-domain orchestration spanning terrestrial MNOs and satellite operators with distinct network management systems, trust domains, and business relationships.

The framework does not assume a single monolithic agent. Instead, multiple specialized agents operate at different levels of the orchestration hierarchy and may be deployed across both TN and NTN domains. Each operator domain (terrestrial MNO, satellite operator) hosts its own set of agents, which coordinate through cross-domain interfaces. These agents are organized at three levels. At the *intent level*, an orchestration agent receives high-level service objectives from applications or human operators; for example, "maintain 4K video session continuity for eVTOL aircraft transitioning from urban TN coverage to satellite NTN during cruise phase" (UC 11.5). At the *policy level*, a policy agent translates intents into concrete orchestration policies: handover trigger thresholds, acceptable QoS degradation bounds during transition, preferred satellite-ground station paths, and fallback strategies. At the *execution level*, domain-specific agents interface with NWDAF for predictions, SHE for compute resource allocation, and the handover module for mobility decisions.

A distinctive challenge for TN-NTN orchestration is the multi-operator nature of the integrated network. A terrestrial MNO and an LEO satellite constellation operator may have roaming-style agreements but operate independent network management domains. The cross-domain agents must: (i) negotiate service continuity guarantees across operator boundaries without exposing proprietary topology or capacity data, (ii) authenticate and authorize actions in both domains, a capability that depends on the standardization of the agent authentication framework discussed as Gap 1 in Section 6.2, and (iii) reconcile potentially conflicting optimization objectives (e.g., the satellite operator may prefer handovers that minimize ISL utilization while the MNO prioritizes UE latency).

Complex NTN scenarios illustrate the need for multi-agent coordination. In the intelligent UAV swarm use case (UC 6.14), a *swarm coordination agent* manages UAV search patterns, a *network selection agent* determines whether each UAV connects via TN or NTN based on coverage and capacity, and an *edge orchestration agent* places AI inference workloads across UAV onboard processors, ground-edge, and space-edge tiers. The agents themselves are distributed across the compute continuum based on their latency and connectivity requirements: the swarm coordination agent runs at the ground-edge or regional edge to meet the 500 ms coordination budget of UC 6.14, the network selection agent executes at the ground-edge where it has direct visibility of both TN and NTN radio conditions, and the edge orchestration agent may run at the regional edge or core cloud where it has a global view of available compute resources. When ground connectivity is unavailable, lightweight agent instances on the space-edge or UAV onboard processors can operate autonomously with reduced decision scope until connectivity is restored.

These agents communicate through standardized protocols and resolve conflicts through a hierarchical consensus mechanism; rather than relying on a single central orchestrator, each operator domain maintains its own coordination agent, and cross-domain conflicts are resolved through negotiation between peer agents at the domain boundary. The AI agent framework must support agent discovery across TN-NTN domains, where a satellite-hosted agent must be discoverable by terrestrial agents and vice versa, requiring extensions to the agent registry with NTN-specific capability descriptors (orbital parameters, coverage area, onboard compute profile).

The hierarchical consensus mechanism is realized as a three-step protocol-level exchange with concrete security and accountability guarantees. Step 1 (*authenticated intent advertisement*): The initiating agent sends a signed intent envelope through the Security Edge Protection Proxy (SEPP)-equivalent inter-PLMN interface of 3GPP TS 33.501 [44], carrying the proposed action, parameters, and a scope token indicating which action classes the originator is authorised to request. Step 2 (*policy-checked acceptance or counter-proposal*): The recipient verifies the scope token against operator policy and the active inter-operator SLA, returning a signed acceptance (with audit action ID), a parameter-bounded counter-proposal, or a typed rejection (*policy-violation*, *insufficient-capacity*, *security-anomaly*). Step 3 (*commit and audit*): Each domain commits locally and writes a tamper-evident log entry indexed by the action ID, enabling dispute resolution without exposing internal topology

to the peer. The convergence cost is bounded by the inter-domain communication and computing durations as $2\,\tau_{\text{TN-NTN}} + 2\,\tau_{\text{compute}} < 200\,\text{ms}$, well within the 1–3 s handover preparation window. The spoofing-protection, fine-grained authorization, and dispute-resolution requirements are some of the gaps that 3GPP TS 33.501 (which defines only message-level roaming security via SEPP) does not yet address for autonomous AI-agent action across domains; the ETSI ZSM reference architecture [45, 46] and ZSM-driven security framework [47] provide the closed-loop security model we adopt as a starting point, with the residual standardization shortfall elaborated as Specification Gap 1 in Section 6.2.

3.5 NKEF for NTN knowledge exposure

NKEF exposes network information to authorized third-party AI applications through standardized APIs, allowing Retrieval-Augmented Generation (RAG) and network-aware AI services [18]. For NTN integration, NKEF must expose five additional knowledge types that do not exist in terrestrial networks:

- *Constellation topology*: Real-time positions and predicted trajectories of all satellites in the serving constellation, including orbital parameters, current phase angles, and predicted pass times over geographic areas of interest. Updated at 1–10 second intervals from ephemeris data.
- *Coverage predictions*: Spatiotemporal maps indicating when satellite coverage is available for a given location, predicted signal quality (RSRP, SINR) based on link budget calculations, and coverage gap durations between successive satellite passes.
- *ISL routing options*: Available inter-satellite paths between any two points in the constellation, per-path latency and capacity estimates, and congestion status. This enables AI agents and third-party applications to make informed decisions about data routing.
- *Ground station status*: Backhaul capacity, current utilization, and predicted load for each ground station, enabling backhaul-aware service placement and handover decisions as demonstrated in [19].
- *Space-edge compute availability*: Available compute resources (processor type, memory, accelerator capacity) on satellites within range, their current utilization, and predicted availability windows, which are essential for AI workload offloading decisions described in Section 3.3.

Third-party AI applications follow a RAG pattern when querying NTN knowledge. Consider a maritime logistics application managing autonomous cargo vessels: The application's LLM agent queries NKEF with "retrieve coverage and ISL routing options for North Atlantic corridor, next 6 hours." NKEF returns predicted satellite pass schedules, expected link quality during each pass, and recommended ground stations for data offload. The LLM synthesizes this into an optimized communication schedule for the vessel fleet, avoiding periods of coverage gaps and routing through uncongested ground stations. This NTN-specific RAG integration extends the terrestrial NKEF capabilities (coverage analytics, QoS catalogs, network analytics) defined in [18] to the spatiotemporal domain unique to satellite networks.

4. HANDOVER MANAGEMENT FRAMEWORK

4.1 Handover scenarios in TN-NTN

Handover in NTN differs fundamentally from terrestrial networks in both the cause and triggering mechanism. In terrestrial NR, the UE is the primary mobility source: As it moves toward a cell edge, RSRP degrades relative to neighboring cells, triggering measurement-report-based handover events (A3/A5). In NTN, the satellite moves while the UE may be stationary: All users within a satellite beam experience nearly uniform RSRP because they are at approximately equal distance from the satellite, eliminating the pronounced signal gradient that terrestrial handover relies on. Consequently, NTN handover shifts from a reactive, UE-measurement-driven paradigm to a proactive, network-controlled paradigm. The network predicts handover timing using satellite ephemeris data (orbital position and velocity vectors broadcast via SIB19), and configures the UE with CHO candidates that include time-based and location-based trigger conditions in addition to supplementary measurement-based conditions [6]. This deterministic predictability of satellite orbits is the key enabler for the AI-driven orchestration proposed in this paper: The NWDAF can compute handover schedules minutes to hours in advance, transforming what would be a reactive signaling storm into an orderly, pre-planned sequence of transitions.

We identify four canonical handover scenarios that the proposed architecture must support. In satellite-to-satellite handover scenarios within LEO constellations, each satellite remains visible from a given ground cell for only 5–10 minutes depending on orbital altitude and minimum elevation angle, making intra-NTN handover the most frequent mobility event in integrated satellite-terrestrial networks. Consequently, intra-orbit handovers occur at intervals on the order of tens of seconds for dense Walker Delta deployments (e.g., 550 km altitude, 53° inclination). Because orbital trajectories are deterministic and published as TLE sets, serving-satellite pass times and Doppler profiles can be predicted hours in advance via SGP4 propagation. The primary challenges are threefold: (i) handover frequency, where at 550 km altitude a UE may experience 40–60 handovers per hour (including both inter-satellite and intra-satellite beam handovers) under

continuous service, imposing substantial signaling load on the access network; (ii) Doppler compensation, where relative velocities of ~7.5 km/s between LEO satellites and ground users produce maximum radial frequency offsets up to ±270 kHz at 28 GHz Ka-band at typical elevation angles, requiring pre-compensation at both the aNB and UE; and (iii) ISL path reconfiguration, where the when the serving satellite changes, inter-satellite routing tables must be updated to maintain end-to-end connectivity through the constellation mesh, a problem that scales with the number of active ISL hops [19]. A Walker Delta ISL topology provides four links per satellite (IntraForward and IntraBackward for intra-plane connectivity, InterLeft and InterRight for adjacent orbital planes).

When a UE transitions from satellite to terrestrial coverage (hand-out), the session must be transferred from the satellite segment to a terrestrial aNB as the UE enters an area with ground-based coverage. This transition involves a shift from satellite propagation characteristics (high delay, large Doppler, wide beamwidth) to terrestrial parameters with sub-millisecond propagation, negligible Doppler, and narrow-cell geometry. The Timing Advance (TA) value must be re-estimated from tens of milliseconds (satellite) to microseconds (terrestrial), and the UE must switch from NTN-specific scheduling request procedures to standard NR random access. Despite these complexities, hand-out is an opportunity: Terrestrial links offer lower latency, higher spectral efficiency, and reduced power consumption at the UE, making timely execution beneficial for both QoS and energy management. The NTN-enhanced NWDAF (Section 3.2) facilitates this transition by predicting terrestrial cell availability along the UE trajectory and pre-computing hand-out timing that minimizes service interruption.

The reverse scenario, terrestrial-to-satellite hand-in, occurs when a UE loses terrestrial coverage, for instance when a vehicle moves from an urban corridor into a rural, maritime, or disaster-affected region, and must be transferred to satellite service to maintain session continuity. The hand-in must prepare the UE for substantially higher propagation delay, reduced link budget margins, and intermittent line-of-sight conditions. Pre-compensation of TA and Doppler before the first NTN random access attempt is essential to avoid repeated preamble failures [6]. Service-level implications include increased jitter and reduced peak throughput; and the orchestration layer must therefore renegotiate QoS parameters and, where possible, pre-position content or migrate edge-application state to the space-edge tier (Section 3.3) before the hand-in completes. The cross-domain AI agent (Section 3.4) coordinates this migration by negotiating resources with the satellite operator domain while maintaining session continuity guarantees established in the terrestrial segment.

Inter-orbit handover involves transitions between different orbital layers (LEO, MEO, and GEO) in multilayer constellations, and is driven primarily by service requirements rather than coverage loss. A delay-sensitive interactive session may prefer LEO (~4.7 ms RTT [19]), whereas a bulk-transfer or broadcast service may tolerate MEO (~270 ms) or GEO (~612 ms) latency in exchange for wider coverage footprints and fewer handovers. Inter-orbit transitions impose the largest parameter discontinuities: TA changes by orders of magnitude, Doppler profiles differ qualitatively, and beam geometry shifts from narrow spot beams (LEO) to continental beams (GEO). The AI-driven orchestration framework described in Section 3 treats orbit selection as an additional degree of freedom in the handover decision space, enabling dynamic tier assignment based on predicted traffic load, satellite visibility windows, and end-to-end path quality.

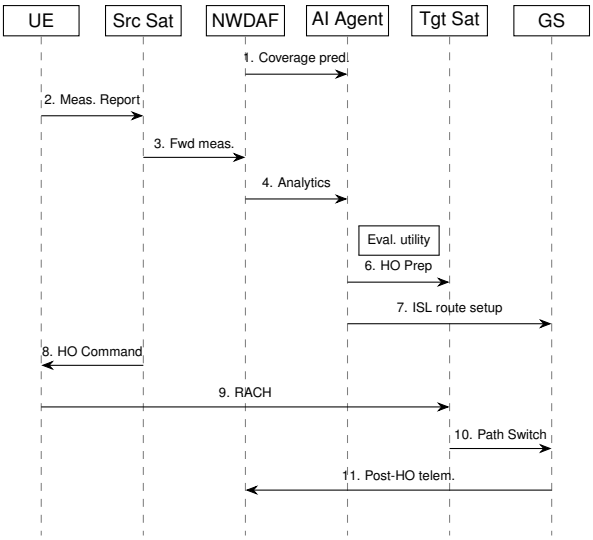

Figure 3 – Satellite-to-satellite handover signaling flow with AI-driven orchestration.

4.2 AI-driven handover decision flow

The handover decision in an AI-native TN-NTN architecture is fundamentally a multi-objective optimization problem whose inputs span radio measurements, orbital mechanics, backhaul state, and service-level requirements. This subsection formalizes the decision flow by mapping it onto the architectural components introduced in Section 3.

The handover decision relies on multiple input data sources aggregated by the NTN-enhanced NWDAF, which collects three categories of input for handover decision support. *Satellite trajectory and coverage predictions* are derived from SGP4 propagation on TLE data to compute per-satellite visibility windows, elevation angles, and Doppler profiles for each ground cell over a configurable prediction horizon (typically 10–30 minutes). *UE measurement reports* provide real-time radio conditions including RSRP, SINR, and Channel Quality Indicator (CQI) for

both the serving and candidate target cells. *Ground-segment state* encompasses per-ground-station backhaul utilization, buffer occupancy, and available ISL routes collected through the telemetry pipeline and exposed through the NKEF. Based on these inputs, the AI agent evaluates candidate handover targets along six decision factors:

1. *Target satellite signal quality*: RSRP and SINR projections computed by combining ephemeris-derived geometry (elevation angle, slant range) with the channel model that accounts for free-space path loss, atmospheric absorption, and stochastic shadowing at 28 GHz (i.e., the Ka band).
2. *Ground station backhaul capacity*: Real-time and predicted utilization of the feeder link and terrestrial backhaul connecting each candidate ground station to the core network. Congested ground stations are penalized to avoid the throughput degradation documented in prior work.
3. *ISL availability and routing cost*: The number of ISL hops, per-hop propagation delay (2–7 ms depending on intra-plane vs. inter-plane spacing), and aggregate link utilization along candidate ISL paths to uncongested ground stations.
4. *Service continuity requirements*: QoS class identifier (5QI), Guaranteed Bit Rate (GBR), and Packet Delay Budget (PDB) associated with the active PDU sessions, which constrain the maximum tolerable handover interruption time.
5. *Energy efficiency*: UE power consumption implications of the candidate target, including Doppler pre-compensation complexity and required transmit power adjustments for the new link budget.
6. *Ping-pong risk*: Estimated remaining dwell time on the candidate satellite, computed from orbital geometry, to avoid immediate re-handover.

The end-to-end handover decision flow follows a four-phase pipeline that integrates these factors into an actionable orchestration sequence. In the *prediction phase*, the NWDAF Analytics Logical Function (AnLF) continuously updates satellite coverage maps, beam load forecasts, and ground-station congestion predictions using time-series models trained on historical telemetry. In the *evaluation phase*, the cross-domain AI agent receives a handover trigger, either a measurement-report-based event (A3/A5) or an ephemeris-based timer, and queries the NWDAF for the current prediction set. The agent formulates the handover as a constrained optimization: Maximize a weighted utility function of signal quality, backhaul capacity, and remaining dwell time subject to the PDB and GBR constraints of all active flows. In the *execution phase*, the agent issues the handover command specifying the target satellite, target ground station, ISL route (if applicable), and handover timing. In the *monitoring phase*, the NWDAF tracks post-handover KPIs (RTT, throughput, packet loss) against pre-handover predictions, enabling closed-loop model refinement through a reward signal fed back to the AI agent's policy network.

Formally, the handover decision space is defined as $\mathcal{A} = \{(s_j, g_k, \mathbf{r}_{jk}) \mid s_j \in \mathcal{S}_{\text{cand}},\ g_k \in \mathcal{G},\ \mathbf{r}_{jk} \in \mathcal{R}_{jk}\}$, where s_j is a candidate target satellite, g_k is a candidate ground station, and $\mathbf{r}_{jk}$ is an ISL route from s_j to g_k. The agent selects the action that maximizes a weighted utility:

$$a^* = \arg\max_{a \in \mathcal{A}} \sum_{i=1}^{6} w_i\, f_i(a) \tag{1}$$

where $f_1(a)$ is the normalized SINR at the target satellite, $f_2(a)$ is the normalized remaining dwell time at the target, $f_3(a)$ is the normalized available backhaul capacity at ground station g_k, $f_4(a)$ is the inverse of the ISL hop count (fewer hops preferred), $f_5(a)$ is the inverse of the estimated handover interruption time, and $f_6(a)$ is an energy efficiency indicator reflecting the satellite power budget impact of the handover path. The weights $w_i \geq 0$, $\sum_i w_i = 1$, are operator-configurable and can be mapped from 5QI profiles to prioritize latency-sensitive or throughput-sensitive services. The optimization is subject to:

$$\text{RTT}(a) \leq \text{PDB}_{\max} \tag{2}$$

$$BW_(a) \geq \text{GBR}_{\min} \tag{3}$$

$$t_{\text{dwell}}(s_j) \geq t_{\text{dwell,min}} \tag{4}$$

$$\text{SINR}(s_j) \geq \text{SINR}_{\text{th}} + \Delta_{\text{hyst}} \tag{5}$$

The problem is combinatorial in nature, with $|\mathcal{A}| = |\mathcal{S}_{\text{cand}}| \times |\mathcal{G}| \times \overline{|\mathcal{R}|}$ candidate actions, where $\overline{|\mathcal{R}|}$ is the average number of distinct ISL routes per satellite–ground-station pair. For small action spaces (e.g., 3–5 visible satellites, 2–3 reachable ground stations), the AI agent performs exhaustive evaluation of *(1)*; for larger constellations, it employs a learned policy network trained via proximal policy optimization on historical handover episodes.

The weights w_i in *(1)* are not free parameters to be hand-tuned per experiment, but are derived systematically from 5QI service-class profiles published in 3GPP TS 23.501, with three calibration steps. First, we partition the service classes into three intent profiles: *latency-sensitive* (5QI 1, 65, 82–85, covering conversational voice, mission-critical push-to-talk, and discrete automation), *throughput-sensitive* (5QI 2, 7–9, covering video and TCP-based traffic), and *reliability-sensitive* (5QI 3, 79–80, covering V2X and low-latency machine-type traffic). Within each profile, the weight on the dominant decision factor is set to 0.45–0.55 and the remaining factors share the residual mass roughly proportional to their normalized standard deviation over recent measurement windows. For example, the latency-sensitive profile uses $(w_1, w_2, w_3, w_4, w_5, w_6) \approx (0.20, 0.10, 0.10, 0.10, 0.45, 0.05)$, emphasizing f_5 (inverse of the handover-interruption-time) over the access-link SINR f_1, whereas the throughput-sensitive profile shifts mass to f_3 (backhaul-capacity), and

the reliability-sensitive profile spreads mass across f_1, f_2, and f_5 to penalise low-SINR, low-dwell-time targets even if their other factors are favorable. Second, the latency-versus-energy trade-off is captured by treating f_5 (inverse of the handover-interruption-time) and f_6 (energy efficiency) as a Pareto front: When both are flagged as "must-respect" by the service profile, the agent applies a lexicographic preference (handover-interruption-time first, then energy among the surviving candidates) rather than collapsing the trade-off into a single scalar; this prevents a low-energy long-interruption candidate from masquerading as optimal under a naive weighted sum. Third, the weights are empirically refined through the closed-loop monitoring described in Section 4.2: Post-handover KPIs (RTT, throughput, packet loss, dwell-time-violations) are scored against the pre-handover predictions, and a slow-loop gradient update adjusts the weights in the direction that reduces the residual KPI error. A full sensitivity analysis on the BrightLight testbed (Section 5.2) under ground-truth handover traces, including comparison against uniform-weight and best-single-factor baselines, is deferred to follow-up experimental work.

The closed-loop pipeline must complete within the handover preparation window T_{prep}, which is constrained by the serving satellite's angular velocity and beam geometry. For a 550 km LEO satellite, T_{prep} is typically 1–3 s before the serving satellite drops below the minimum elevation angle. The pipeline stages have the following approximate latency budget: SGP4-based constellation prediction, ~10–50 ms (precomputed and updated periodically); NWDAF analytics query for ground-station load forecast, ~20–100 ms; AI agent utility computation over the candidate set $\mathcal{A}$, ~5–50 ms depending on $|\mathcal{A}|$; and handover command transmission to the satellite, ~3.3 ms per ISL hop plus ~1.7 ms ground-to-LEO one-way propagation. The total pipeline latency is therefore ~40–205 ms, well within the 1–3 s preparation window. Because the entire decision pipeline executes at the ground-based NWDAF and AI agent, terrestrial compute resources are used for prediction and optimization; the satellite receives only the final handover execution command, keeping onboard processing requirements minimal.

This four-phase pipeline distinguishes the proposed architecture from prior DRL-based handover schemes [15, 27, 28] in three respects: (i) The decision agent is embedded within the standardized NWDAF framework rather than operating as a standalone module, enabling reuse of existing 3GPP data collection and analytics interfaces; (ii) ground-segment backhaul state is explicitly incorporated as a first-class decision factor rather than being abstracted away; and (iii) the closed-loop monitoring phase enables online model adaptation without requiring offline retraining.

Beyond these three structural differences, our framework also subsumes the input features used by prior DRL handover schemes (RSRP/Doppler/dwell in [15]; graph topology in [27, 37]; LSTM-predicted beam load in [28]) and additionally ingests ground-station backhaul utilization, ISL routing cost, and 5QI-derived service constraints. The action space is enlarged from target-satellite selection to the joint $(s_j, g_k, \mathbf{r}_{jk})$ triple. Direct numeric comparison against the prior reported gains (64% handover-frequency reduction in [15], 23% throughput improvement in [28], 86.7% RTT reduction in [19]) is unreliable because each study uses a different scenario, which is precisely the motivation for the unified-benchmarking proposal in Section 5.3; a head-to-head experiment under identical scenarios is scoped in the validation roadmap.

4.3 Backhaul-aware path selection

Conventional satellite handover algorithms optimize for access-link quality, specifically RSRP, SINR, and remaining satellite visibility, while treating the ground segment as an unconstrained resource. This assumption breaks down in practice: Ground stations serving dense LEO constellations experience time-varying load as satellites sweep overhead, and terrestrial backhaul links connecting ground stations to the core network are subject to congestion, particularly in remote or underserved regions where satellite connectivity is most needed [19].

To formalize the backhaul congestion problem, consider a UE served by satellite s_1 through ground station GS_1 with backhaul capacity C_1. When the aggregate traffic through GS_1 exceeds C_1, queuing delay at the ground segment dominates the end-to-end RTT, causing the total delay to oscillate between 440 and 660 ms as TCP's congestion avoidance interacts with the bottleneck buffer, a pattern characteristic of bufferbloat [19]. A conventional handover triggered at time t selects a target satellite s_2 based on radio-link criteria but routes traffic through the same congested GS_1 (or another equally congested station), yielding no RTT improvement. The fundamental issue is that the handover decision space does not include the backhaul dimension.

To address this limitation, the proposed backhaul-aware path selection extends the handover decision space to include ISL-assisted rerouting to alternative ground stations with sufficient spare capacity. When the NWDAF detects that the serving ground station's backhaul utilization exceeds a configurable threshold (e.g., 80% of C_1), the AI agent evaluates ISL paths from the target satellite to alternative ground stations with sufficient spare capacity. In a Walker Delta topology (e.g., 550 km altitude, 53° inclination), each satellite maintains four ISL connections (IntraForward and IntraBackward within the same orbital plane, and InterLeft and InterRight to adjacent planes), enabling multi-hop routing across the constellation mesh [19]. The agent selects the ISL route

that minimizes the weighted sum of propagation delay (computed from inter-satellite distance at the speed of light, ~3.3 ms per hop) and backhaul congestion at the destination ground station.

The backhaul-aware handover mechanism was experimentally validated [19] using a controlled LEO first-shell topology (550 km, 53° inclination) with three satellites, two ground stations, and a single UE generating 20 Mbps TCP offered load. In the baseline scenario (Scenario A), handover at t=30 s transitions the UE to a new satellite but maintains the path through a congested ground station (GS_1, 10 Mbps backhaul capacity). Post-handover RTT remains in the 440–660 ms oscillation band, and throughput stabilizes at ~9.3 Mbps, well below the offered load. In the ISL bypass scenario (Scenario B), the same handover reroutes traffic via ISL to a neighboring satellite connected to an uncongested ground station (GS_2, 50 Mbps backhaul). The RTT drops from ~720 ms to a steady-state of ~55 ms within seconds, an 86.7% reduction, consistent with the ISL path's expected round-trip propagation delay of ~50 ms plus scheduling overhead. Throughput initially spikes to ~48 Mbps as TCP buffers drain, then settles to ~20 Mbps matching the offered load. These results demonstrate that satellite switching alone does not resolve ground-segment congestion, and that backhaul-aware path selection is a critical design dimension for future 6G Space-Terrestrial Integrated Network (STIN) handover architectures.

Realizing backhaul-aware path selection in practice requires tight coordination among the three architectural components defined in Section 3. The *NTN-enhanced NWDAF* continuously monitors ground-station backhaul utilization through the telemetry pipeline and generates congestion predictions using time-series forecasting. The *cross-domain AI agent* receives these predictions alongside satellite coverage forecasts and formulates the joint satellite-and-ground-station selection as the constrained optimization described in Section 4.2. The *SHE space-edge tier* provides the ISL routing substrate: Onboard routing agents at each satellite maintain neighbor-state tables and respond to rerouting commands from the ground-based agent. This three-component coordination loop executes within the handover preparation window (typically 1–3 s before the predicted serving-satellite pass boundary), ensuring that the ISL path is established before the handover completes. The NKEF exposes the resulting backhaul analytics, including ground-station load distributions, ISL path latency maps, and rerouting event logs, to external consumers such as network slice managers and third-party application orchestrators.

4.4 Handover optimization challenges

While the AI-driven handover architecture presented in the preceding subsections addresses the core decision-making and path-selection problems, several optimization challenges remain open and warrant discussion. A key challenge is ping-pong avoidance in LEO constellations, where the high orbital velocity of satellites (~7.5 km/s) creates scenarios in which a UE near the boundary between two satellite footprints may oscillate rapidly between serving cells. Unlike terrestrial ping-pong, which is mitigated by hysteresis margins of a few dB, LEO ping-pong involves rapidly changing geometry: The elevation angle to each candidate satellite changes by several degrees per second. Ephemeris-based prediction can identify boundary-zone UEs and extend the dwell-time constraint in the handover utility function, but the prediction accuracy degrades when atmospheric scintillation or multipath from terrain features perturbs the received signal. The NWDAF's closed-loop monitoring (Section 4.2) can detect ping-pong patterns in real time and dynamically adjust the hysteresis margin, but the optimal adaptation rate, which must balance responsiveness against stability, remains an open research question.

3GPP Release 17 introduced DAPS handover to reduce interruption time by maintaining simultaneous connections to source and target cells, and extending this mechanism to TN-NTN transitions introduces two specific complications. First, the propagation delay difference between the TN and NTN legs can exceed 50 ms, complicating the PDCP reordering window and potentially causing unnecessary retransmissions. Second, the UE must simultaneously maintain two timing advance values and two Doppler compensation loops, increasing baseband processing complexity and power consumption. The space-edge SHE tier could offload part of this complexity by performing server-side PDCP aggregation, but this requires regenerative payload capability and incurs additional onboard processing latency.

LEO satellites typically illuminate the ground through multiple spot beams steered electronically or mechanically, and the resulting beam management problem introduces another layer of handover complexity. Intra-satellite beam handover (beam switching) occurs more frequently than inter-satellite handover and must be coordinated with the inter-satellite handover decision to avoid cascading transitions. Testbed platforms supporting 28 GHz beamforming with phased-array antennas can enable experimental evaluation of beam-switching dynamics, but the interaction between beam management and ISL rerouting introduces a combinatorial expansion of the decision space that current DRL formulations [15] do not address.

Frequent LEO handovers also generate substantial control-

plane signaling overhead: Each inter-satellite handover involves measurement reports, handover commands, random access procedures, and path-switch messages to the core network. At constellation scale, with tens of thousands of UEs across hundreds of satellites, the aggregate signaling load can stress the AMF/SMF capacity. Two mitigation strategies align with our architecture. First, CHO with ephemeris-assisted preconfiguration reduces the per-handover signaling exchange by eliminating the measurement-report-trigger-command round trip. Second, the NWDAF can aggregate handover decisions for groups of UEs within the same beam, issuing batch handover commands that amortize the signaling cost across multiple sessions.

Quantifying the signaling-vs.-latency trade-off: The backhaul-aware extension adds two messages on the inter-satellite control plane (ISL-route-setup and ISL-route-confirm) atop the six conventional NR handover messages. Because these messages traverse the satellite operator's ground control segment rather than the AMF/SMF, the increase is $\approx$ 33% on the satellite-side plane but only the **12% additional signaling overhead** measured empirically in [19] on the cellular core plane. Against the 86.7% RTT reduction (Section 4.3), the cost is $\approx$ 0.14% signaling per percent of RTT improvement. For the 1 584-satellite Walker-Delta benchmark constellation at 550 km / 53° (Section 5.3), with ~32 beams per satellite and 10^4 UEs per beam, the aggregate handover rate is $\sim 1.1 \times 10^6$ handovers/s; conditional-handover and batched-decision mitigations (batching factor 8–16) reduce the core-plane signalling load to ~4–8 $\times 10^5$ messages/s, within the capacity envelope of modern cloud-native 5G cores. The dominant scale-out bottleneck is the satellite-telemetry ingest at the NWDAF rather than the cellular control plane, mitigated by hierarchical per-satellite pre-aggregation before forwarding to the central NWDAF AnLF.

When the TN-NTN handover involves different operators (e.g., a terrestrial MNO and a satellite operator), training a unified handover prediction model requires data from both domains, motivating the use of federated learning for cross-operator model development. Direct data sharing may violate regulatory or commercial confidentiality constraints. Federated learning [31, 48] offers a privacy-preserving alternative: Each operator trains a local model on its domain-specific data (terrestrial radio measurements, satellite telemetry) and shares only gradient updates with a central aggregator hosted by a trusted third party or within the NWDAF. The non-i.i.d. nature of satellite versus terrestrial data distributions poses convergence challenges that heterogeneous federated optimization methods such as FedProx [32] can partially address, but the intermittent connectivity between satellite and ground aggregation points introduces additional synchronization barriers that are unique to the NTN context.

5. TESTBED EVALUATION

5.1 Testbed comparison

Validating the proposed AI-native orchestration architecture requires testbed platforms that simultaneously support realistic satellite dynamics, protocol-level network emulation, and AI analytics integration. No single existing platform satisfies all three requirements. To contextualize our validation approach using the BrightLight platform, we present a systematic comparison of eight representative NTN testbed and simulation platforms spanning emulation, simulation, and hardware-in-the-loop paradigms.

Table 2 compares the platforms across eleven dimensions critical to NTN orchestration research. The comparison reveals several structural observations. The most fundamental distinction among the compared platforms is the emulation versus simulation trade-off. Platforms divide into two families. Pure simulators (Hypatia, Celestial, spacecraft-ns3) offer detailed physical-layer and orbital modeling but cannot execute real protocol stacks in real time, limiting their utility for end-to-end QoS validation and AI model serving experiments. Emulation-based platforms (OpenSand, OAI 5G-NTN) run real protocol stacks but lack constellation-scale orbital dynamics. BrightLight bridges this gap through its hybrid architecture: Linux network namespaces provide per-entity isolated kernel network stacks for real packet processing, while the optional NS-3 mmWave module supplies detailed 28 GHz channel modeling with beamforming, free-space path loss, atmospheric absorption, and stochastic shadowing. The eBPF-accelerated packet path achieves throughput exceeding 100 000 packets per second per link with minimal CPU overhead [19].

Regarding 5G core integration, among the compared platforms only BrightLight and OAI 5G-NTN integrate a production-grade 5G core network. BrightLight couples Open5GS (providing AMF, SMF, UPF, and related control-plane functions) with UESIM/gNB for functional attach and data paths, enabling end-to-end registration, PDU session setup, and GTP-U forwarding over emulated satellite links [19]. This integration is essential for validating NTN-aware session management and QoS procedures proposed in 3GPP Release 20.

Inter-satellite link support is another critical differentiator, as ISL emulation is essential for backhaul-aware handover validation. BrightLight implements ISL with four links per satellite following the Walker Delta pattern, with propagation delays computed from Haversine inter-node distances at the speed of light (2–7 ms depending on intra-plane vs. inter-plane spacing) [19]. The topology module supports configurable polar exclusion (e.g., $> 53°$ latitude for inclined-orbit constellations). Hypa-

Table 2 – Extended testbed comparison matrix for NTN platforms

Platform	Language	Architecture	Isolation	Wireless Sim	5G Core	Max Sats	Real-time	ISL	Monitoring	Open Src
BrightLight [19]	Rust	Hybrid	Namespaces	NS-3 mmWave	Open5GS	2000+	Yes	Yes	Prometheus/ Grafana	Yes
OpenSand	C/C++	Emulation	VMs	DVB-S2/RCS2	No	Limited	Yes	Limited	GUI	Yes
Hypatia	Python/ C++	Simulation	Virtual	ns-3	No	1000+	No	Yes	Log files	Yes
Celestial	Go	Simulation	Containers	Basic	No	100s	No	Yes	Log files	Yes
spacecraft-ns3	C++	Simulation	Virtual	ns-3 advanced	No	1000+	No	Yes	ns-3 trace	Yes
ESA 6G-LINO [38]	—	Hardware	Physical sat.	Real RF	OAI aNB	Limited	Yes	No	HW monitor	No
OAI 5G-NTN [49]	C	Software	Containers	Real NR stack	Partial	Limited	Possible	No	OAI tools	Yes
Space-O-RAN [23]	—	Conceptual	—	O-RAN NTN	Partial	—	—	Planned	—	Partial

tia and spacecraft-ns3 also model ISLs but within their discrete-event simulation frameworks, precluding real-time protocol interaction during ISL rerouting events.

For hardware-in-the-loop validation, ESA's 6G-LINO project [38] is the only platform with a physical satellite (CubeSat) and real RF link, providing ground-truth validation of NTN channel models. However, its single-satellite configuration and lack of ISL support limit its applicability to constellation-scale or handover-oriented experiments. OpenAirInterface 5G-NTN [49] provides a complete NR protocol stack implementation suitable for NTN-specific PHY/MAC validation but does not integrate orbital mechanics or constellation management.

From an architectural perspective, Space-O-RAN [23] proposes an O-RAN-aligned architecture for NTN with distributed RICs at ground and satellite segments. While architecturally innovative, it remains at the conceptual/ simulation stage and does not yet provide an executable testbed. BrightLight's three-tier design, comprising a kernel-accurate OS substrate, composable simulator core, and lightweight UI, yields a reproducible, instrumented STIN testbed that is straightforward to extend with multi-satellite constellations or alternative RAN stacks while retaining architectural fidelity to 3GPP NTN and 5GS specifications [19].

Finally, production-grade monitoring and observability differentiate BrightLight from all other compared platforms. Its native Prometheus exporter publishes per-node CPU utilization, per-link latency, packet loss ratios, throughput, and handover events to Grafana dashboards with configurable alerting [19]. This telemetry pipeline directly maps to the NWDAF data collection architecture proposed in Section 3.2, enabling seamless integration of testbed metrics with AI analytics stubs. Alternative platforms offer limited monitoring: OpenSand provides a GUI with basic link statistics, Hypatia and Celestial rely on post-experiment log analysis, and spacecraft-ns3 generates ns-3 trace files requiring offline processing.

5.2 BrightLight for orchestration validation

The BrightLight testbed provides the experimental foundation for validating the proposed AI-driven orchestration architecture. This subsection describes how BrightLight's architectural components map to the orchestration functions defined in Section 3 and summarizes the published validation results that establish the testbed's credibility for NTN orchestration research.

BrightLight's modular architecture maps naturally to the orchestration components defined in Section 3 through four integration points [19]. The Prometheus exporter collects per-node CPU utilization, per-link latency, packet loss, throughput, and handover events at sub-second granularity, conforming to the OpenMetrics exposition model and enabling direct ingestion by an NWDAF AnLF stub for time-series forecasting of handover timing and ground-station load. The simulation manager controls workload injection, impairment profiles, and handover triggering through Unix socket IPC, providing a programmatic interface for AI agent-initiated handover execution and topology reconfiguration without simulation restart. The orbital calculator implements SGP4 propagation on TLE data, computing line-of-sight distances and propagation delays at each time step to generate the satellite trajectory predictions that feed the NWDAF's coverage analytics. Finally, YAML-based configuration templates, where YAML stands for "YAML Ain't Markup Language", support three representative LEO constellation scales, namely a mega-constellation (~10 000 satellites at 550 km, 53° inclination), a medium constellation (~3 000 satellites across multiple orbital shells), and a smaller constellation (~650 satellites in polar orbits), enabling comparative evaluation of handover frequency, ISL topology density, and ground-station access patterns.

Three sets of published validation results establish BrightLight's suitability for NTN orchestration research [19]: End-to-end RTT measurements across the three orbital regimes, specifically LEO (~4.7 ms at 550 km altitude),

MEO (~270 ms at 8,000–20,000 km), and GEO (~612 ms at 35,786 km), aligned with the theoretical predictions derived from propagation path length and the speed of light, confirming accurate delay modeling across heterogeneous network segments. Scalability evaluation with constellation sizes ranging from 100 to 2 000 satellites demonstrated sub-linear CPU scaling (from ~0.02% to below 0.005% per satellite beyond 1 000 nodes) and memory usage of approximately 50 MB per 100 entities, enabling handover algorithm evaluation at realistic constellation densities. The backhaul-aware handover experiments, detailed in Section 4.3, validated that ISL-assisted rerouting to uncongested ground stations significantly reduces post-handover RTT and eliminates throughput bottlenecks.

Reproducibility is ensured because all experiments use deterministic simulation manager timing, YAML-based constellation configuration, and pcap-validated packet flows through inter-namespace traffic captures [19]. The BrightLight framework is publicly available as an open-source project, enabling independent reproduction of the reported results and extension with additional orchestration components.

5.3 Testbed gaps and recommendations

The testbed comparison in Section 5.1 reveals that no existing platform, including BrightLight, fully integrates AI orchestration logic with NTN emulation at constellation scale. We identify five capability gaps and formulate corresponding recommendations for next-generation STIN testbeds.

The first gap concerns native NWDAF and AI analytics integration. Current NTN testbeds treat analytics as a post-experiment activity: Metrics are collected during execution and analyzed offline. None integrates a real-time NWDAF instance that consumes telemetry, generates predictions, and feeds decisions back into the network control loop during the experiment. Future testbeds should embed an NWDAF stub with configurable AnLF modules that ingest OpenMetrics streams and expose analytics via standardized NWDAF service interfaces.

The second gap involves AI model serving and migration across compute tiers. Evaluating the space-edge SHE tier (Section 3.3) requires the ability to deploy, monitor, and migrate AI models between emulated satellite payloads (space-edge), ground-station co-located servers (ground-edge), and cloud instances. No current testbed supports AI model lifecycle management across these tiers. Testbed architectures should include a model registry service, per-tier inference runtime environments, and migration protocols that account for ISL bandwidth constraints and contact-window limitations.

Third, multi-operator federation support is absent from current testbeds. The cross-domain AI agent (Section 3.4) operates across terrestrial MNO and satellite operator boundaries. Testing this capability requires testbed support for multiple administrative domains with independent policy engines, authentication mechanisms, and data sovereignty constraints. Testbed platforms should support namespace-level or container-level domain isolation with configurable inter-domain APIs, enabling experiments in federated learning, cross-operator handover negotiation, and data-sharing agreements under differential privacy [50] constraints.

The fourth gap relates to digital twin synchronization [51]. Several use cases in TR 22.870 (e.g., UC 11.7 for airspace management) rely on network digital twins that maintain a real-time replica of the physical network state. Validating digital-twin-based orchestration requires the testbed to simultaneously maintain a physical emulation layer and a synchronized digital representation that AI agents can query and manipulate. The digital twin should consume the same telemetry as the NWDAF and provide a queryable graph database of network topology, resource utilization, and predicted state trajectories.

The fifth gap is the absence of standardized benchmarking suites. Comparing NTN testbed capabilities is currently hampered by the absence of standardized benchmark scenarios. Each platform reports results under different constellation configurations, traffic models, and handover trigger conditions, making cross-platform comparison unreliable [52]. The research community should develop a benchmark suite comprising: (i) a reference LEO constellation (e.g., 1 584 satellites at 550 km, 53° inclination) with standardized ground-station placement; (ii) canonical traffic profiles (CBR, TCP bulk transfer, HTTP adaptive streaming) with defined offered loads; (iii) handover scenario templates covering the four types described in Section 4.1; and (iv) KPI definitions (handover interruption time, ping-pong rate, throughput recovery time, signaling overhead per handover) with measurement methodology specifications.

These five gaps collectively define the research agenda for next-generation NTN testbed development. Addressing them would enable end-to-end validation of AI-driven orchestration architectures, from NWDAF prediction through AI agent decision to SHE-tier execution, over realistic, reproducible, and comparable experimental conditions.

Finally, to make the comparison in Table 2 reproducible rather than narrative, we further propose a unified benchmarking methodology rooted in three principles that the comparison rows already approximate. First, *configuration invariants*: all reported results should be tagged with the constellation altitude, inclination, satellite count, ground-station count and placement, ISL topol-

ogy (Walker Delta, Walker Star, or grid), and UE distribution model, so that headline numbers can be reproduced from the metadata. Second, *traffic-and-trigger contracts*: each KPI is reported under one of three canonical traffic loads (low: 1 Mbps CBR; medium: 20 Mbps TCP; high: HTTP/3 adaptive video at 25 Mbps) and one of three handover trigger sets (A3-event, ephemeris-time-based, and the AI-driven utility of *(1)*), yielding a 3×3 comparison matrix that exposes whether a platform's reported gain is driven by traffic mix, trigger choice, or genuine algorithmic improvement. Third, *measurement provenance*: each KPI carries a measurement-point annotation (UE-side, gNB-side, ground-segment, end-to-end) and a counting rule (e.g., "handover-success-rate counted as fraction of triggered handovers that complete within 100 ms of the predicted commit time"), preventing the cross-paper ambiguity where "handover latency" is variously defined. Adopting this scheme requires no new instrumentation beyond what Section 5.2 already describes for BrightLight, and the resulting benchmarking specification can be hosted as an open testbed-and-results manifest analogous to the MLPerf model-card pattern.

6. STANDARDIZATION ALIGNMENT AND OPEN CHALLENGES

6.1 3GPP Release 19–20 NTN progress

The integration of non-terrestrial networks into the 3GPP ecosystem has progressed through a series of releases, each expanding the scope and maturity of NTN support. We summarize the trajectory from Release 17 through the anticipated Release 21 timeline and map our proposed contributions to this evolving landscape.

Releases 17 and 18 established the baseline NTN capability within 3GPP. Release 17 introduced foundational NR-NTN support for transparent satellite payloads, including timing advance enhancements, Doppler precompensation, and ephemeris-assisted cell selection. Release 18 extended coverage to NB-IoT and eMTC over NTN, enabling direct-to-satellite IoT connectivity for asset tracking and environmental monitoring.

Release 19, frozen in December 2025 [7], represents a significant maturation of NTN capabilities. Key work items include: (i) *NTN for NR Phase 3*, which adds support for regenerative payloads where the satellite hosts a full gNB stack, Ka/Ku-band operation, and enhanced HARQ procedures for high-latency links; (ii) *IoT-NTN enhancements* for store-and-forward operation and discontinuous coverage scenarios; (iii) *inter-RAT mobility* between NR-NTN and E-UTRA-NTN, enabling legacy device migration; and (iv) initial study items on *NTN-assisted positioning* leveraging LEO ranging signals. Notably, Release 19 does not yet address AI/ML-based mobility optimization for NTN or NWDAF extensions for satellite telemetry ingestion, which are precisely the areas targeted by our architecture.

The ongoing Release 20 [8] introduces the "Satellite and NTN in 5G Architecture" work item, which addresses the core network implications of deep NTN integration. Stage 1 requirements were frozen in June 2025 and include: (i) Architectural support for satellite-terrestrial session continuity, including AMF/SMF extensions for NTN-aware session management; (ii) QoS framework enhancements to accommodate the variable and orbit-dependent latency profiles of LEO, MEO, and GEO links; and (iii) network slicing for NTN, enabling differentiated service treatment for IoT, broadband, and safety-critical NTN traffic. In parallel, Release 20 advances the NWDAF specification (TS 23.288) with new analytics IDs and data collection enhancements, though satellite-specific data sources such as ephemeris, ISL topology, onboard resource utilization are not yet incorporated. The 6G use case study in TR 22.870 [9] was completed in Release 20, providing the requirements baseline referenced throughout this paper.

Looking ahead, Release 21, expected in the 2027–2028 timeframe, is anticipated to deliver the first normative 6G specifications, building on the ITU-R IMT-2030 framework [1] and the use case requirements established in TR 22.870. ETSI's vision document on NTN in 6G [33] anticipates that Release 21 will formalize: (i) AI-native network management functions with NTN awareness; (ii) multi-orbit constellation management within a unified 6G core; and (iii) standardized interfaces for space-edge computing resource exposure.

The proposed contributions in this paper are designed to align with this release trajectory. Our proposed NTN-enhanced NWDAF (Section 3.2) and cross-domain AI agent (Section 3.4) are designed to be *forward-compatible* with the Release 19–20 NWDAF framework while introducing the satellite-specific extensions that Release 21 is expected to standardize. Specifically, the ephemeris-based beam prediction, ISL topology analytics, and space-edge resource monitoring functions proposed in this paper can be realized as new AnLF modules within the existing NWDAF decomposition, while requiring only the addition of new analytics IDs and output schemas to the existing Nnwdaf_AnalyticsInfo and Nnwdaf_AnalyticsSubscription service operations. The backhaul-aware handover mechanism validated experimentally [19] provides empirical evidence for the feasibility of these extensions ahead of their formal standardization.

6.2 Specification gaps for AI-driven NTN orchestration

Building on the four specification gaps described in [18] for AI-native 6G architectures, we identify five gaps that are specific to, or significantly amplified by, the NTN dimension. These gaps represent areas where current 3GPP normative texts and ongoing work items do not yet provide the mechanisms required to realize the AI-driven TN-NTN orchestration envisioned in this paper.

The first and most impactful gap concerns cross-domain agent authentication and authorization. Our prior work in [18] identified the absence of a standardized authentication and authorization framework for AI agents operating across network function boundaries as a fundamental gap affecting 15 agent-class use cases in TR 22.870. In the NTN context, this gap is substantially amplified: An AI agent orchestrating handover or resource allocation must interact with network functions belonging to *different administrative domains*, such as a terrestrial MNO, a satellite operator, and potentially a HAPS provider, each with distinct trust boundaries, security policies, and regulatory jurisdictions. Current 3GPP security specifications (TS 33.501) define inter-PLMN security for roaming but do not address the fine-grained, per-action authorization semantics required for autonomous AI agents that issue configuration commands (e.g., beam steering, ISL rerouting) across these domains in real time.

The second gap relates to AI model lifecycle management for space-edge deployments. No current specification addresses the full lifecycle (deployment, versioning, monitoring, update, and decommissioning) of AI models hosted on satellite payloads. The challenges are unique to the space segment: (i) Model updates must be transmitted during narrow ground-station contact windows with integrity and authenticity guarantees; (ii) rollback mechanisms are needed in case an updated model degrades onboard inference quality, but satellite storage is limited; (iii) performance monitoring of onboard models requires uplink telemetry that competes with user-plane traffic for bandwidth. Extending the O-RAN AI/ML workflow framework or the NWDAF Model Training Logical Function (MTLF) to encompass space-edge nodes would be a natural starting point, but neither currently accounts for the intermittent connectivity and constrained resources of the satellite environment.

A third gap arises in federated learning across satellite-terrestrial operator boundaries. Federated learning is a compelling paradigm for training AI models across distributed satellite and ground nodes without centralizing raw data [31, 48]. However, when FL spans multiple operator domains, as required for joint TN-NTN model training, three unsolved standardization issues arise: (i) *Privacy guarantees*: Differential privacy [50] or secure aggregation protocols must be mandated when gradient updates traverse inter-operator interfaces, yet no 3GPP specification defines the required privacy budget parameters or aggregation trust model; (ii) *convergence under non-i.i.d. data*: Satellites in polar vs. equatorial orbits observe fundamentally different traffic and channel distributions, and while algorithmic solutions such as FedProx [32] exist, the signaling framework for communicating data distribution metadata to the aggregator is absent; (iii) *Byzantine robustness*: A compromised satellite node could inject poisoned model updates, and current specifications lack mechanisms for verifiable gradient provenance or outlier detection at the aggregation point.

The fourth gap concerns semantic communication for satellite telemetry and analytics. The bandwidth constraints of satellite links make them a natural candidate for semantic communication [53], in which only the *meaning* of data, rather than its raw bit-level representation, is transmitted. In the context of NTN orchestration, two applications are particularly impactful: (i) *Compressed model/analytics exchange*, where NWDAF analytics outputs (e.g., beam load predictions, handover recommendations) are encoded semantically before transmission over capacity-constrained ISLs or feeder links; and (ii) *satellite telemetry compression*, where onboard sensors generate high-volume health and environmental data that must be distilled to decision-relevant summaries before downlink. However, no current 3GPP specification defines semantic encoding formats, semantic-aware QoS classes, or the interfaces through which a semantic encoder/decoder pair is negotiated between communicating network functions.

To strengthen this point with concrete operational mechanics, we describe how the cross-domain AI Agent prioritizes *semantic features* over *raw telemetry* when bandwidth on the feeder or ISL path is scarce. The agent maintains a per-stream relevance score $\rho_i \in [0, 1]$ that captures the decision-influence of stream i on the current set of active orchestration intents (e.g., handover preparation, beam-load balancing). At inference time, instead of forwarding the raw telemetry tensor $\mathbf{x}_i$, the agent forwards a learned compact embedding $\phi_\theta(\mathbf{x}_i)$ whose dimensionality is reduced from raw size $|\mathbf{x}_i|$ to a target $k_i = \lceil \rho_i \cdot k_{\max} \rceil$ semantic dimensions. Recent semantic-IoT designs achieve up to 99% reduction in communication volume and 80% reduction in end-to-end transmission delay relative to raw bit-level transmission under matched accuracy targets [54], and cross-modal graph semantic encoders demonstrate that high-quality reconstruction is possible from a small number of attention-weighted semantic kernels [53]. Within the AI agent, this translates into three policy rules: (a) For low-relevance streams (e.g., periodic housekeeping telemetry from a satellite that is not in the candidate set of the next handover), the agent transmits only a heartbeat plus aggregated statistics; (b) for handover-relevant streams (RSRP, Doppler, beam-load), it transmits the semantic embedding extracted by

the per-stream encoder; and (c) for safety-critical streams (anomaly indicators, link-failure precursors), it falls back to lossless raw-bit transmission. The negotiation of which encoder/decoder pair to use across the ISL is precisely the missing 3GPP interface noted in this gap; we anticipate that a future analytics-ID extension in TS 23.288 will need to carry the semantic-encoder identifier alongside the analytics output schema, so that the consumer NF can decode the embedding correctly.

The fifth and final gap is the need for NWDAF extensions to support NTN-specific data sources. The current NWDAF specification (TS 23.288 [20]) defines data collection from terrestrial network functions (AMF, SMF, PCF, OAM) but does not include satellite-specific data sources. For effective NTN orchestration, the NWDAF must ingest: (i) *Satellite ephemeris data* to predict beam coverage windows, handover timing, and ground-station contact schedules; (ii) *ISL topology and link-state information* to evaluate backhaul path quality for handover decisions; (iii) *onboard resource telemetry* (CPU utilization, thermal state, power budget) from space-edge computing nodes; and (iv) *atmospheric and weather data* that affect Ka/Ku-band link budgets. Defining new data collection interfaces, analogous to the existing Naf, Namf, and Nsmf service-based interfaces, for these NTN-specific sources is a prerequisite for any AI-driven NTN optimization. We note that the analytics IDs and output formats would also require extension: For example, a "Satellite Beam Load Prediction" analytics ID that outputs per-beam load forecasts indexed by orbital position and time would enable proactive handover and resource reservation.

6.3 Limitations and validation roadmap

The architecture proposed in this paper is conceptual and standards-aligned; quantitative validation of several claims is scoped as follow-on testbed work on the BrightLight platform (Section 5). Six future-work items define the testbed validation roadmap: (R1) sweep ephemeris-update interval and TLE age to calibrate the timing-error-to-handover-failure curve refining the drift-detection rules of Section 3.2; (R2) benchmark a representative AI-pipeline workload mix against Algorithm 1 placements with exhaustive-search ground truth on small constellations; (R3) run the DRL agents of [15, 27] and our utility-based agent head-to-head on the 1 584-satellite Walker-Delta scenario under the unified-benchmarking methodology of Section 5.3; (R4) extend the open-loop latency bound of Section 4.2 to a discrete-time control-theoretic stability analysis with Lyapunov-style bounded-error guarantees under bursts of visibility-transition; (R5) publish reference results under the canonical 3×3 scenario matrix as an open YAML testbed-and-results manifest alongside the BrightLight release; and (R6) develop the Markov-decision-process sketch of the orchestration loop into a full constrained-MDP specification enabling direct comparison with prior resource-allocation frameworks. Each item is reachable on the existing testbed without new hardware.

7. CONCLUSIONS AND FUTURE DIRECTIONS

This paper presented an AI-driven orchestration architecture for integrated satellite-terrestrial 6G networks, bridging the gap between AI-driven network intelligence and non-terrestrial network management. The key contributions and findings are summarized as follows.

First, we extended the seven-component AI-native 6G architecture from [18] with four NTN-specific modules: An NTN-enhanced NWDAF that ingests satellite telemetry and ephemeris data for predictive handover analytics, a space-edge computing tier within SHE enabling onboard satellite AI inference, a cross-domain AI agent for multi-operator TN-NTN orchestration, and an NKEF extension exposing constellation topology and coverage predictions. The architecture is organized into three orchestration domains (terrestrial, non-terrestrial, and a cross-domain orchestration layer) operating on principles of predictive intelligence, backhaul-aware management, and zero-touch operation.

Second, the AI model placement framework demonstrated that the four-tier compute continuum (space-edge, ground-edge, regional edge, core cloud) requires latency-aware, power-aware, and connectivity-aware placement decisions that fundamentally differ from terrestrial-only scenarios. The dynamic model migration capability addresses the unique challenge of maintaining AI service continuity as LEO satellites traverse their orbits.

Third, the handover management architecture addresses four TN-NTN transition types with AI-driven decision flows that consider satellite trajectory predictions, ground station backhaul congestion, and ISL routing alternatives. Building on our experimental validation of backhaul-aware ISL-assisted rerouting [19], the architecture integrates NWDAF predictions to enable proactive, rather than reactive, handover decisions.

Fourth, the comparative evaluation of eight NTN testbed platforms revealed that no existing testbed fully integrates AI orchestration logic with realistic NTN emulation. Our testbed framework [19] uniquely combines real protocol stacks, constellation-scale support (2,000+ satellites), ISL connectivity, and production-grade monitoring, providing the most comprehensive platform for validating the proposed architecture.

Five specification gaps were identified that must be ad-

dressed in 3GPP Releases 20–21 to realize AI-driven NTN orchestration: cross-domain agent authentication, AI model lifecycle management for space-edge, federated learning across satellite-terrestrial operators, semantic communication for satellite telemetry, and NWDAF extensions for NTN analytics. In parallel, five testbed capability gaps were listed, namely native NWDAF integration, AI model serving across compute tiers, multi-operator federation, digital twin synchronization, and standardized benchmarking suites, which collectively define the research agenda for next-generation NTN validation platforms.

Future work will pursue five directions: (i) Full implementation and validation of the NTN-enhanced NWDAF and AI agent within our testbed framework, enabling closed-loop orchestration experiments; (ii) multi-UE scenarios with network slicing across TN-NTN domains to evaluate orchestration scalability; (iii) federated learning for handover models trained across multiple operators without exposing proprietary network topology; (iv) Doppler-aware frequency management integrated with the handover decision engine; and (v) digital twin integration for predictive constellation management with what-if analysis capabilities. The proposed architecture provides a systematic foundation for transforming 6G NTN management from fragmented, rule-based approaches to unified, AI-driven orchestration.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their constructive feedback. Murat Parlakisik conducted this research in a personal capacity, independent of the author's affiliation with Amazon. The opinions and findings expressed are solely those of the author and do not represent the views of Amazon.

REFERENCES

[1] International Telecommunication Union. *Framework and overall objectives of the future development of IMT for 2030 and beyond*. Tech. rep. Recommendation ITU-R M.2160-0. International Telecommunication Union Radiocommunication Sector, 2023.

[2] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang. "The roadmap to 6G: AI empowered wireless networks". In: *IEEE Communications Magazine* 57.8 (2019), pp. 84–90. DOI: 10.1109/MCOM.2019.1900271.

[3] W. Saad, M. Bennis, and M. Chen. "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems". In: *IEEE Network* 34.3 (2020), pp. 134–142. DOI: 10.1109/MNET.001.1900287.

[4] M. Giordani and M. Zorzi. "Non-terrestrial networks in the 6G era: Challenges and opportunities". In: *IEEE Network* 35.2 (2021), pp. 244–251. DOI: 10.1109/MNET.011.2000493.

[5] P. Pillai, S. Kota, and G. Giambene. "Satellite–5G integration: A network perspective". In: *IEEE Network* 32.5 (2018), pp. 25–31. DOI: 10.1109/MNET.2018.1800037.

[6] 3GPP. *Study on New Radio (NR) to support non-terrestrial networks*. Tech. rep. TR 38.811, V15.4.0. 3rd Generation Partnership Project, 2020.

[7] 3GPP. *5G non-terrestrial networks in Release 19*. Tech. rep. Technical Summary. 3rd Generation Partnership Project, 2025.

[8] 3GPP. *Release 20 overview and stage 1 freeze*. Tech. rep. Technical Summary. 3rd Generation Partnership Project, 2025.

[9] 3GPP. *Study on 6G use cases and related potential requirements*. Tech. rep. TR 22.870, V1.0.1, Release 20. 3rd Generation Partnership Project, 2025.

[10] G. Svistunov, A. Akhtarshenas, D. Lopez-Perez, M. Giordani, G. Geraci, and H. Yanikomeroglu. *Bridging Earth and space: A survey on HAPS for non-terrestrial networks*. arXiv preprint arXiv:2510.19731. 2025.

[11] O. Abbasi, A. Yadav, H. Yanikomeroglu, N. D. Dao, G. Senarath, and P. Zhu. "HAPS for 6G networks: Potential use cases, open challenges, and possible solutions". In: *IEEE Wireless Communications* 31 (2024), pp. 324–331.

[12] M. Mahbub, M. M. Saym, S. Jahan, et al. "A holistic survey of UAV-assisted wireless communications in the transition from 5G to 6G: State-of-the-art innovations, challenges, and opportunities". In: *J. Netw. Comput. Appl.* 237 (2025), p. 104131.

[13] J. Chen, H. Zhang, and Z. Xie. *Space-air-ground integrated network (SAGIN): A survey*. arXiv preprint arXiv:2307.14697. 2023.

[14] F. Rinaldi, A. Iera, and G. Araniti. "Toward 6G non-terrestrial networks". In: *IEEE Network* 36.1 (2022), pp. 113–120. DOI: 10.1109/MNET.011.2100191.

[15] X. Liu et al. "Satellite network handover scheme based on improved deep reinforcement learning". In: *Proc. Int. Conf. Space Inf. Netw.* Springer, 2024.

[16] K. Aldubaikhy. "A low-complexity hybrid handover strategy for LEO NTN: Balancing stability and link quality". In: *Sensors* 26.5 (2026), p. 1449. DOI: 10.3390/s26051449.

[17] S. Javaid, R. A. Khalil, N. Saeed, B. He, and M. Alouini. "Leveraging large language models for integrated satellite-aerial-terrestrial networks: Recent advances and future directions". In: *IEEE Open J. Commun. Soc.* 6 (2025).

[18] M. Parlakisik and N. Guney. "A unified AI-native 6G architecture for 3GPP AI use cases: Systematic categorization and identification of gaps in specifications". In: *Proc. IEEE Int. Conf. Commun. (ICC)*. 2026.

[19] M. Parlakisik and E. Ozturk. "A modular testbed for integrated satellite–terrestrial networks". In: *Proc. 8th Int. Conf. Adv. Commun. Technol. Netw. (CommNet)*. Rabat, Morocco, 2025, pp. 1–7. DOI: 10.1109/CommNet68224.2025.11288807.

[20] 3GPP. *Architecture enhancements for 5G system (5GS) to support network data analytics services*. Tech. rep. TS 23.288, V18.5.0. 3rd Generation Partnership Project, 2023.

[21] P. K. Gkonis, N. Nomikos, P. Trakadas, L. Sarakis, et al. "Leveraging network data analytics function and machine learning for data collection, resource optimization, security and privacy in 6G networks". In: *IEEE Access* 12 (2024), pp. 21320–21336. DOI: 10.1109/ACCESS.2024.3359992.

[22] K. Koufos, K. El Haloui, M. Dianati, M. Higgins, J. Elmirghani, M. A. Imran, and R. Tafazolli. "Trends in intelligent communication systems: Review of standards, major research projects, and identification of research gaps". In: *J. Sens. Actuator Netw.* 10.4 (2021), p. 60. DOI: 10.3390/jsan10040060.

[23] E. Baena, P. Testolina, M. Polese, D. Koutsonikolas, J. Jornet, and T. Melodia. *Space-O-RAN: Enabling intelligent, open, and interoperable non-terrestrial networks in 6G*. arXiv preprint arXiv:2502.15936. 2025.

[24] M. H. Rahman, M. S. Hossen, N. H. Stephenson, et al. *A demonstration of self-adaptive jamming attack detection in AI/ML-integrated O-RAN*. arXiv preprint arXiv:2510.09706. 2025. DOI: 10.48550/arXiv.2510.09706.

[25] F. Michel, M. Trevisan, D. Giordano, and O. Bonaventure. "A first look at Starlink performance". In: *Proc. 22nd ACM Internet Measurement Conf. (IMC)*. 2022, pp. 130–136. DOI: 10.1145/3517745.3561416.

[26] F. Muro, E. Baena, T. de Cola, et al. *A 5G NTN emulation platform for VNF orchestration: Design, development, and evaluation*. Authorea Preprint. 2024. DOI: 10.22541/au.170628474.46948465/v1.

[27] M. Chen et al. "Graph-based handover for future aeronautical 6G integrated terrestrial and non-terrestrial networks". In: *Proc. IEEE/AIAA Digital Avionics Systems Conf. (DASC)*. 2024.

[28] D. Fan, S. Zhou, J. Luo, Z. Yang, and M. Zeng. "Joint traffic prediction and handover design for LEO satellite networks with LSTM and attention-enhanced Rainbow DQN". In: *Electronics* 14.15 (2025), p. 3040. DOI: 10.3390/electronics14153040.

[29] O. Kodheli, E. Lagunas, N. Maturo, et al. "Satellite communications in the new space era: A survey and future challenges". In: *IEEE Communications Surveys & Tutorials* 23.1 (2021), pp. 70–109. DOI: 10.1109/COMST.2020.3028247.

[30] Y. Chen et al. "Integrating terrestrial and non-terrestrial networks in 6G: A review of architectures, AI-driven techniques, and sustainability strategies". In: *IEEE Access* 12 (2024), pp. 1–20.

[31] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. "Communication-efficient learning of deep networks from decentralized data". In: *Proc. AISTATS*. 2017, pp. 1273–1282.

[32] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. "Federated optimization in heterogeneous networks". In: *Proc. MLSys*. 2020.

[33] ETSI. *Vision on non-terrestrial networks in 6G system*. White Paper. 2024.

[34] E. Coronado, R. Behravesh, T. Subramanya, A. Fernandez-Fernandez, M. S. Siddiqui, X. Costa-Perez, and R. Riggio. "Zero touch management: A survey of network automation solutions for 5G and 6G networks". In: *IEEE Communications Surveys & Tutorials* 24.4 (2022), pp. 2535–2578.

[35] H. Daniel, O. Alhussein, C. Li, J. Liang, and E. Damiani. *LLM-enabled NWDAF: A step toward AI-native 6G network intelligence*. Research Square Preprint. 2026.

[36] X. Hu, H. Song, S. Liu, and W. Wang. "Velocity-aware handover prediction in LEO satellite communication networks". In: *Int. J. Satellite Commun. Netw.* 36.6 (2018), pp. 451–459. DOI: 10.1002/sat.1250.

[37] H. Yu, W. Gao, K. Zhang, L. Ma, and W. Sun. "A graph reinforcement learning-based handover strategy for LEO satellites under power grid scenarios". In: *Aerospace* 11.7 (2024), p. 511. DOI: 10.3390/aerospace11070511.

[38] Fraunhofer IIS. *6G LINO: LEO satellite testing infrastructure for 6G*. ESA Project. 2025.

[39] NASA JPL. *Towards space edge computing and onboard AI for real-time Earth observation*. IEEE LEO Satellites Report. 2024.

[40] Y. Shi, J. Zhu, C. Jiang, L. Kuang, and K. B. Letaief. *Satellite edge artificial intelligence with large models: Architectures and technologies*. arXiv preprint arXiv:2504.01676. 2025.

[41] Q. Chen, Z. Wang, X. Chen, J. Wen, D. Zhou, S. Ji, M. Sheng, and K. Huang. "Space–ground fluid AI for 6G edge intelligence". In: *Engineering* (2025).

[42] H. Zhou, C. Hu, Y. Yuan, et al. "Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities". In: *IEEE Communications Surveys & Tutorials* 27.3 (2025), pp. 1955–2005.

[43] R. Zhang, H. Du, Y. Liu, D. Niyato, J. Kang, Z. Xiong, A. Jamalipour, and D. I. Kim. "Generative AI agents with large language models for satellite networks". In: *IEEE J. Sel. Areas Commun.* 42.12 (2024), pp. 3581–3596.

[44] 3GPP. *Security architecture and procedures for 5G system*. Tech. rep. TS 33.501, V18.x. 3rd Generation Partnership Project, 2024.

[45] ETSI. *Zero-touch network and service management (ZSM); reference architecture*. Tech. rep. GS ZSM 002, V1.1.1. European Telecommunications Standards Institute, 2019.

[46] C. Benzaïd and T. Taleb. "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions". In: *IEEE Network* 34.2 (2020), pp. 186–194. DOI: 10.1109/MNET.001.1900252.

[47] G. Chollon, D. Ayed, R. Asensio Garriga, et al. "ETSI ZSM-driven security management in future networks". In: *Proc. IEEE Future Networks World Forum (FNWF)*. 2022, pp. 334–339. DOI: 10.1109/FNWF55208.2022.00065.

[48] S. Niknam, H. S. Dhillon, and J. H. Reed. "Federated learning for wireless communications: Motivation, opportunities, and challenges". In: *IEEE Communications Magazine* 58.6 (2020), pp. 46–51.

[49] S. Kumar, A. K. Meshram, A. Astro, et al. "OpenAirInterface as a platform for 5G-NTN research and experimentation". In: *Proc. IEEE Future Networks World Forum (FNWF)*. 2022, pp. 500–506. DOI: 10.1109/FNWF55208.2022.00094.

[50] C. Dwork and A. Roth. "The algorithmic foundations of differential privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014), pp. 211–487. DOI: 10.1561/0400000042.

[51] B. Mao, Y. Zhou, Y. Liu, and N. Kato. "Digital twin satellite networks toward 6G: Motivations, challenges, and future perspectives". In: *IEEE Network* 38.1 (2024), pp. 54–60.

[52] W. Jiang, Y. Zhan, X. Xiao, and G. Sha. "Network simulators for satellite-terrestrial integrated networks: A survey". In: *IEEE Access* 11 (2023), pp. 98269–98292.

[53] H. Xie, Z. Qin, G. Y. Li, and B.-H. F. Juang. "Deep learning enabled semantic communication systems". In: *IEEE Trans. Signal Processing* 69 (2021), pp. 2663–2675. DOI: 10.1109/TSP.2021.3071210.

[54] Y. Qi. "Computationally efficient approach for 6G-AI-IoT network slicing and error-free transmission". In: *Int. J. Netw. Manag.* 35.2 (2025). DOI: 10.1002/nem.70007.

AUTHORS

Murat Parlakisik is currently a senior engineer at Amazon, Seattle, WA, USA, working on large-scale cloud and networking infrastructure, and is concurrently pursuing his Ph.D. degree at the University of North Dakota. He received his M.S. degree in computer science in 2001 and later completed a machine learning specialization at the Georgia Institute of Technology. He is the inventor of multiple issued patents covering SDN and radio access network control, and has authored publications on QoS-enabled OpenFlow, control-plane performance management, and integrated satellite–terrestrial network testbeds. His current research interests include 6G networks, non-terrestrial networks, AI-native network architecture design, AI-driven orchestration, and satellite–terrestrial integration.

Ertan Ozturk received his M.S. and Ph.D. degrees in electrical and computer engineering from the Illinois Institute of Technology (IIT), Chicago. He previously worked at Motorola as a senior systems engineer and later served for 14 years as a faculty member in the Department of Electrical and Electronics Engineering at Bülent Ecevit University. He also led embedded systems development activities at ERETNA Technology. Since August 2022, he has been with the University of North Dakota. His research interests include wireless communications, communication theory, and integrated satellite–terrestrial networks for next-generation wireless systems. He has authored/co-authored about 50 peer-reviewed publications and currently leads research on AI-enabled satellite–terrestrial networking technologies, including a NASA EPSCoR-supported project.

Nazli Güney is currently an R&D projects manager at TT Mobil, the mobile network operator of the Turk Telekom Group. Dr. Güney received her B.S. (with high honors), M.S. and PhD degrees in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2001, 2003, and 2009, respectively, and was a research assistant during the years 2001-2009 at the same institution. Following her PhD studies, she joined the telecommunications industry as a researcher, and she worked for several renowned telecom companies in Turkey. Her interest for innovation management and new idea / product development activities led her to receive an Executive MBA degree in 2014, also from Bogazici University. Dr. Güney continues to pursue an academic career alongside her professional activities, and she is with Istanbul Rumeli University as an assistant professor of computer engineering. With an ambition to contribute to enabling novel applications and services for 5G and beyond systems, her research interests lie within the broad category of wireless communications system design.