UE-Density-Aware Multipath Management for System Utility Enhancement of 5G-6G Dualsteer

Sohei Itahara, Akito Suzuki, Takeo Ogawara, and Masaki Suzuki
KDDI Research, Inc., Japan
so-itahara@kddi.com

***Abstract*—Dualsteer (DS) is a core-network (CN)-level multi-generation aggregation method that leverages multipath transport protocols, including multipath QUIC, to integrate legacy and new radio access network (RAN) without modifying the legacy RAN. This RAN transparency enables smooth, cost-effective network migration. Our flow-level radio resource allocation modeling reveals the possibility of DS in performance degradation due to RAN transparency. If DS persists in activating both paths at high user equipment (UE) density, the system utility is inferior to that of single-path methods. Furthermore, the packet-level simulations reveal that existing multipath transport mechanisms cannot avoid this risk. Even though existing mechanisms reduce the traffic on the weaker path, they still waste radio resources due to persistent splitting and keep both paths active. To address the problem without introducing RAN impact, we propose a UE-density-aware DS that dynamically controls splitting and switching using only information available on the CN or the UE. Large-scale simulations with 12 BSs and 100 UEs demonstrate that the proposed scheme improves application-level goodput at high UE density by 21.5% over existing schedulers, while retaining bandwidth-aggregation benefits at low UE density.**

***Index Terms*—Dualsteer, Network migration, Multipath, Dual connectivity, MPQUIC**

I. Introduction

For the successful and sustainable realization of 6G, network migration has attracted greater attention than in previous generations, including 4G and 5G [1], [2]. The mobile network operators (MNOs) already operate mature 4G and 5G systems and expect 6G to evolve seamlessly on top of legacy infrastructure while reusing existing facilities as much as possible [2]. A fundamental enabler of such migration is multi-generation traffic aggregation [1], [3]–[6], which allows a user equipment (UE) to communicate simultaneously with legacy and new base stations (BSs) and improves throughput, spectrum efficiency, and reliability. In the 5G introduction, EUTRAN-NR dual connectivity (EN-DC), also known as non-standalone (NSA), has been widely commercialized [7].

A key lesson from DC deployment is sustainability, since DC often requires extensive updates to legacy BSs solely for compatibility even when existing hardware remains otherwise adequate [5]. Motivated by sustainability, dualsteer (DS), a core network (CN)-level aggregation method [1], [4]–[6], has emerged as an alternative where traffic aggregation is handled in the CN without coordination between new and legacy BSs by using layer 4 (L4) multipath protocols such as multipath TCP (MPTCP) and multipath QUIC (MPQUIC). In DS, the traffic aggregator is the user plane function (UPF) in CN [8]. Since BSs remain agnostic to aggregation and CN functions are already virtualized and centralized, DS can reduce operational effort and enhance long-term sustainability [5]. A high-level comparison of DC and DS is depicted in Fig. 1.

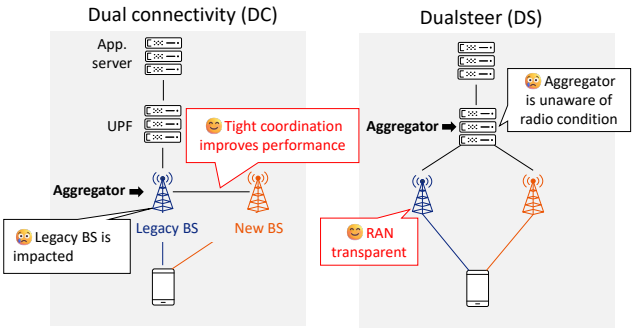

Fig. 1. High-level summary of multi-generation aggregation methods. In DC, BS aggregates the traffic. In DS, the UPF in CN aggregates traffic.

This paper investigates the degradation of the system utility caused by DS under high UE density. Existing transport multipath studies mainly focus on per-UE performance [9]–[11] and fairness to single-path UEs [12], [13], leaving system-level utility unclear. In contrast, the DC literature [14] has extensively studied system utility and reports that the utility gain from multi-connectivity diminishes as UE density increases, approaching that of single-cell selection, but never falling below it. This raises a concern for DS in high-density environments, because DS is inherently RAN-unaware and does not control the RAN. The UPF cannot observe radio conditions such as signal-to-interference and noise ratio (SINR), which can lead to inefficient traffic steering and reduced system utility [5].

Therefore, we have a research question: *whether RAN-transparent DS becomes inferior to single-generation selection in terms of system utility under high UE density.* To answer this question, we conduct flow-level radio resource allocation modeling and packet-level simulations. The results reveal that DS with existing L4 multipath mechanisms can be inferior to simple single-generation (SG) selection under high UE density in terms of system utility. In particular, the flow-level evaluation shows that if DS persists in splitting, meaning both paths remain active, system utility degrades not only compared to DC but also compared to simple SG under high UE density,

ISBN 978-3-903176-82-9

as described in Sec. III. The packet-level evaluation further shows that existing multipath scheduling algorithms [9]–[11] and congestion control algorithms (CCAs) [12], [13], [15] tend to maintain persistent splitting, resulting in utility degradation, as described in Sec. IV.

Aiming to mitigate degradation caused by excessive splitting while retaining bandwidth-aggregation benefits without introducing RAN impact, we propose UE-density-aware DS, called UE-DA DS. UE-DA DS uses cell reselection information in the system information block (SIB) to determine path priority and suppresses splitting based on the number of active sessions in the cell, steering traffic to the priority path at high UE density. Large-scale simulations with 12 BSs and 100 UEs show that UE-DA DS achieves 21.5% higher application-level goodput than existing multipath methods at high UE density, while retaining bandwidth-aggregation benefits at low UE density.

The contributions of this paper are as follows:

- We formulate a flow-level radio resource allocation model for DS and SG by extending the DC formulation in [14]. Using this model, we demonstrate that persistent dual-path activation in DS degrades system utility at high UE density, making DS inferior to local SG.
- We conduct packet-level simulations with four L4 multipath schedulers [9]–[11] and three CCAs [12], [13], [15]. The results show that existing mechanisms cannot avoid persistent splitting, which wastes radio resources on inefficient paths and reduces system utility.
- We propose UE-density-aware DS, called UE-DA DS, which switches between single and splitting modes based on an estimated per-cell load using only UE and UPF information. UE-DA DS preserves RAN transparency and improves application-level goodput at high UE density while retaining bandwidth-aggregation benefits at low UE density.

To the best of our knowledge, [13] is the only L4 multipath study that explicitly addresses system efficiency degradation under high UE density. It targets wireless local area network (WLAN)-cellular MPTCP and mitigates contention-induced throughput loss by suppressing the secondary path using queue occupancies of access points. In contrast, we analyze DS-specific degradation driven by spectrum efficiency (SE) degradation in cellular-cellular multipath and develop a RAN-transparent method relying solely on UE and UPF information. Moreover, we show that secondary-path suppression [13] alone is insufficient to prevent persistent splitting and the associated utility loss in DS, which is detailed in Sec. IV.

II. Related Works

Dual connectivity. DC [3] is a traffic aggregation method commercialized as EN-DC [3] during the transition from 4G to 5G, also known as the NSA architecture. DC enables the UE to simultaneously connect to both new and legacy BSs. Traffic is distributed and terminated at the legacy BS and UE, respectively. Based on the interaction between legacy and new BS, the legacy BS distributes the traffic to the new BS.

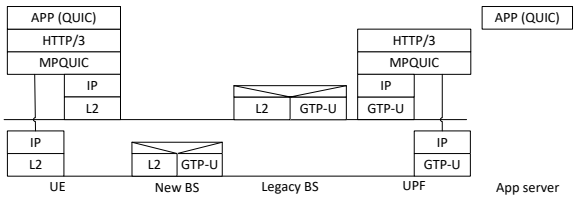

Fig. 2. Protocol Stack of DS using MPQUIC.

Thanks to tight coordination between legacy and new BSs, e.g., frequent sharing of SINR, resource block (RB) allocation, buffer sizes, and loads [16], DC enables improvements in reliability [17] and throughput [14].

Beyond the performance benefits of DC, a key lesson learned from NSA deployments is that the tight coordination required between BSs in DC significantly impacts the legacy RAN, often necessitating hardware replacement or substantial upgrades of existing BSs [1], [5]. Given that the RAN infrastructure has accumulated across multiple generations, such replacements are not cost-effective. Therefore, a multi-generation aggregation approach that requires updates only to the CN, while leaving the legacy RAN unchanged, is highly desirable. This is because the CN is already largely virtualized, whereas the RAN remains predominantly purpose-built [5], at the time of 6G introduction. Therefore, this paper focuses on DS [1], [4]–[6], an emerging cost-efficient alternative to DC.

Dualsteer. DS [1], [4]–[6], [8], [18] is a multi-generation aggregation method by CN without impacting RAN. Traffic is split and aggregated at the UPF and UE, with transparency to the RAN. In more detail, the UE and the UPF support MPQUIC proxy [19] and establish two QUIC connections between the UE and the UPF via an HTTP/3 proxy architecture. Application flows are encapsulated and tunneled over HTTP/3 using the appropriate CONNECT methods [20]. The protocol stack is illustrated in Fig. 2.

DS is based on L4 multipath technology with WLAN and cellular, e.g., MPQUIC [19] and MPTCP [21], which is extensively investigated. In the L4 multipath protocol, there are two major areas of research: scheduling and CCA. The multipath scheduling algorithm [9], [10] aims to improve the throughput while avoiding the head-of-line (HoL) blocking. Coupled CCA [12] manages the congestion window (CWND) of each path in a mutually coordinated manner, optimising overall throughput while maintaining fairness for single-path flows. For both areas, a cross-layer control approach with the lower layer is proposed. Based on the MAC queue length of the WLAN AP, CCA [13] is proposed to suppress the usage of weak paths, and a scheduler [11] controls the packet.

Unlike the integration of WLAN and cellular networks, this paper focuses on cellular-cellular integration [1], [4]–[6]. [4] evaluates the multipath scheduler in DS. [5] compares the DS and DC, clarifying the performance gap. [6] proposes a cross-layer scheduler to improve the latency. Unlike existing DS studies, this paper focuses on the system utility of DS in high UE density environments.

System utility problem in high UE-density. In the DC

literature, a limitation of DC [14], [17], [22] regarding system efficiency is reported. In [22], although increasing the number of simultaneously connected cells per UE improves SE, the multi-cell gain diminishes as the number of cells increases, especially beyond 4. [17] reports that trade-off between reliability and system utility. In [14], the system utility gain of DC over load-aware ideal single-cell allocation diminishes at high UE density. These papers [14], [17], [22] address in DC literature, where an aggregator, i.e., a legacy BS, has detailed radio information, including RB allocation, SINR, and cell load, and control on RAN

Unlike these papers [14], [17], [22] in DC literature, this paper focuses on the DS, in which an aggregator, i.e., the UPF, has no control over the RAN and is unaware of the RAN. For example, the cell to which a UE connects is selected independently in each RAN generation, without coordination between RANs, and the UPF lacks SINR and RB allocation information. We found a critical problem caused by the RAN-unawareness in DS, which is discussed in Sec. III and IV.

III. Radio Resource Allocation Modeling

This section evaluates the impact of RAN unawareness on system utility in the DS using a flow-level radio resource allocation model. We consider two DS variants to evaluate the effect of RAN unawareness at the UPF. Ideal DS serves as an upper bound, assuming the UPF has full RAN knowledge and can optimally steer the traffic, namely legacy-only, new-only, or both. Fully active (FA) DS captures a practical risk case in which the UPF lacks RAN information, and the multipath mechanism keeps both paths persistently active, e.g., with a round-robin scheduler under sufficient traffic. For comparison, we also consider DC and two single-generation baselines. The modeling of DC is followed by [14]. Ideal SG assumes global information and optimizes UE association, whereas local SG selects the generation with the largest SINR with an offset.

To evaluate the system utility, we focus on downlink communication with a sufficiently large application buffer, similar to [14]. This paper focuses on downlink communication, since downlink traffic accounts for 92% of traffic in commercial networks.

A. System model

Fig. 3 shows the system model consists of N UEs, B^{L} legacy BSs, B^{N} new BSs, an UPF, and an application server. The UPF accommodates all BSs and UEs. The legacy and new BSs are operated on different frequencies. For example, considering 5G and 6G, the 5G and 6G BSs operate in frequency range (FR) 1, i.e., below 6 GHz, and in FR3, i.e., 6 GHz to 15 GHz [23].

B. Modeling of system utility of DC, DS, and SG

Dual connectivity: The system utility modeling of DC follows [14]. Let $\mathcal{U}_b$ denote the set of UEs served by BS b. A UE can connect to at most one legacy BS and at most one new BS. Let $\mathcal{B}_u^{\mathrm{L}}$ and $\mathcal{B}_u^{\mathrm{N}}$ denote the candidate sets of legacy and new BSs for UE u, respectively. Given SINR $r_{b,u}$ from

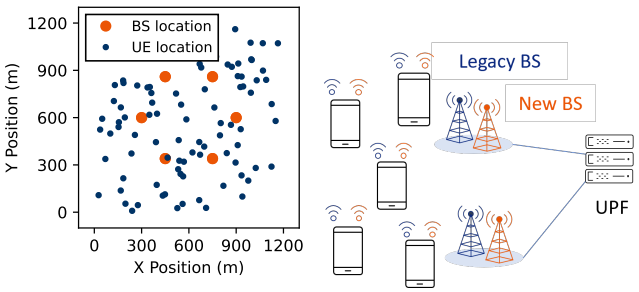

Fig. 3. Example of 100 UE locations and 6 BS locations. At the BS location, new BS and legacy BS are colocated, resulting in 12 BSs.

BS b to UE u, total bandwidth W_b at BS b, and set of all BS $\mathcal{B}$, the throughput of UE u is

$$t_u^{\mathrm{DC}} = \sum_{b \in \mathcal{B}_u^{\mathrm{L}} \cup \mathcal{B}_u^{\mathrm{N}}} w_{b,u} \mu_{b,u}, \tag{1}$$

where $\mu_{b,u} = \log_2(1 + r_{b,u})$, subject to

$$\forall b \in \mathcal{B}, \sum_{u \in \mathcal{U}_b} w_{b,u} \leq W_b, \tag{2}$$

$$\forall u \in \mathcal{U}, \sum_{b \in \mathcal{B}_u^{\mathrm{L}}} \mathbb{1}\{w_{b,u} > 0\} \leq 1, \tag{3}$$

$$\forall u \in \mathcal{U}, \sum_{b \in \mathcal{B}_u^{\mathrm{N}}} \mathbb{1}\{w_{b,u} > 0\} \leq 1. \tag{4}$$

For shorthand, we denote the B-dimensional vector $\boldsymbol{W}$ as $(W_1, \ldots, W_B)$ and the $U \times B$ matrix $\boldsymbol{\mu}$ as $\boldsymbol{\mu}[b, u] = \mu_{b,u}$.

The proportional fair (PF) metrics [24], which balance the throughput and fairness, is used as a system utility in this paper. The system utility of DC L^{DC} is given by the optimal value of

$$\max_{\boldsymbol{w}} \frac{1}{N} \sum_{u \in \mathcal{U}} \log\left(t_u^{\mathrm{DC}}\right) \quad \text{s.t. (2), (3), (4).} \tag{5}$$

Dualsteer: The critical difference between DS and DC is that, in DS, the UPF is RAN-unaware and does not control the RAN. Cell selection and RB allocation are optimized locally within each generation, without coordination across generations. The UPF only controls whether to forward traffic to each generation, relying on transport-layer metrics such as RTT rather than RAN metrics such as SINR and load. To reflect this constraint, we formulate DS as follows.

Let b_u^{L} and b_u^{N} denote the serving cells of UE u in the legacy and new RANs, respectively. The serving cells b_u^{L} and b_u^{N} are determined locally in the legacy and new RANs, which are denoted in Sec III-C2. For each UE u, we introduce binary variables α_u^{L} and α_u^{N} indicating whether the UPF forwards traffic to the legacy and new RANs. For notational convenience, we write α_u^g for $g \in \{\mathrm{L}, \mathrm{N}\}$ to represent α_u^{L} and α_u^{N}, respectively. We denote $\boldsymbol{\alpha}^{\mathrm{L}} = (\alpha_1^{\mathrm{L}}, \ldots, \alpha_N^{\mathrm{L}})$ and $\boldsymbol{\alpha}^{\mathrm{N}} = (\alpha_1^{\mathrm{N}}, \ldots, \alpha_N^{\mathrm{N}})$.

Within each cell b, radio resources are shared equally among active UEs [24], namely UEs with $\alpha_u^g = 1$ on that generation. Let

$$k_b^g = \sum_{u \in \mathcal{U}} \alpha_u^g \mathbb{1}\{b_u^g = b\} \tag{6}$$

denote the number of active UEs in cell b of generation g. When $\alpha_u^g = 1$, UE u receives bandwidth $W_{b_u^g}/k_{b_u^g}^g$ from generation g. We assume $k_{b_u^g}^g \geq 1$ whenever $\alpha_u^g = 1$. The resulting UE throughput is

$$t_u^{\mathrm{CN}} = \sum_{g \in \{\mathrm{L},\mathrm{N}\}} \alpha_u^g \frac{W_{b_u^g}}{k_{b_u^g}^g} \mu_{b_u^g, u}. \tag{7}$$

The system utility of DS L^{DS} is formulated as

$$\frac{1}{N} \sum_{u \in \mathcal{U}} \log\left(t_u^{\mathrm{CN}}\right). \tag{8}$$

To evaluate the impact of RAN-unawareness, we consider two DS variants. Ideal DS serves as an upper bound, where the UPF is assumed to know $\boldsymbol{\mu}$ and optimally selects $\boldsymbol{\alpha}^{\mathrm{L}}$ and $\boldsymbol{\alpha}^{\mathrm{N}}$ as

$$\max_{\boldsymbol{\alpha}^{\mathrm{L}}, \boldsymbol{\alpha}^{\mathrm{N}}} L^{\mathrm{DS}} \tag{9}$$

FA-DS captures persistent dual-path activation without load-aware steering, namely $\boldsymbol{\alpha}^{\mathrm{L}} = \boldsymbol{\alpha}^{\mathrm{N}} = \mathbf{1}$. Note that packet-level simulations in Sec. IV show that existing multipath mechanisms result in persistent dual-path activity.

Single generation allocation: In SG, cell association and RB allocation are optimized locally within each generation, as in DS, and the UE uses only one generation at a time. Using $k_{b_u^g}^g$ in (6), the throughput of UE u and system utility are

$$t_u^{\mathrm{SG}} = \sum_{g \in \{\mathrm{L},\mathrm{N}\}} \alpha_u^g \frac{W_{b_u^g}}{k_{b_u^g}^g} \mu_{u, b_u^g}, \tag{10}$$

$$L^{\mathrm{SG}} = \frac{1}{N} \sum_{u \in \mathcal{U}} \log\left(t_u^{\mathrm{SG}}\right) \tag{11}$$

where $\alpha_u^{\mathrm{L}} = 1 - \alpha_u^{\mathrm{N}}$.

Ideal SG serves as an upper bound where a centralized optimizer has global RAN information and selects $\boldsymbol{\alpha}^{\mathrm{L}}$ to maximize system utility, formulated as

$$\max_{\boldsymbol{\alpha}^{\mathrm{L}}} L^{\mathrm{SG}}. \tag{12}$$

In local SG, the UE selects the generation based on SINR with an offset, similar to inter-frequency selection with q-offset [25]. Given SINR values r^{L} and r^{N} for the legacy and new candidates and offset q, the UE selects the legacy generation if $r^{\mathrm{L}} - q > r^{\mathrm{N}}$. The offset is typically configured by self-organized RAN (SON) to improve system-level performance [26].

C. Setup

We consider a 1,200 m $\times$ 1,200 m area with 6 BS sites and N UEs as described in Fig. 3. Each BS site has a legacy BS and a new BS because the new BS is assumed to be installed at the existing site alongside the legacy BS. The inter-site distance is 300 m. We assume the UEs are fixed as [14], where the positions of UEs are generated by the Poisson point process. The legacy and new BSs operate at 2.0 GHz and 6.0 GHz, respectively, with 10 MHz bandwidth.

1) Path loss modeling: Given the frequency f and distance between BS and UE d, we model the path loss using a shadowing model by

$$20 \log_{10}(4\pi f/c) + 10\, a \log_{10}(d) + X_\sigma, \tag{13}$$

where c is the speed of light, a is the path loss exponent, X_σ is a log-normal shadowing term drawn from $\mathcal{N}(0, \sigma^2)$ in dB. We set $c = 3.0 \times 10^8$ m/s, $a = 3.5$, $\sigma = 8$ dB, and $\mathrm{NF} = 5$ dB, following 3GPP channel modeling [27]. The noise figure of 5 dB is used.

Following [14], we ignore inter-cell interference by assuming ideal interference coordination [28]. We model SINR as

$$r_{b,u} = \frac{p_{\mathrm{tx}}\, l_{b,u}}{N_0 W_b}, \tag{14}$$

where p_{tx} is the BS transmit power, $l_{b,u}$ is the pathloss between BS b and UE u including noise figure, and N_0 is the noise power spectral density.

2) Serving cell selection in each generation: In DS and SG, the serving cell in each generation b_u^g is optimized independently in each generation g by the RAN and does not depend on the UPF traffic-forwarding variables α_u^g. Let

$$K_b^g = \sum_{u \in \mathcal{U}} \mathbb{1}\{b_u^g = b\} \tag{15}$$

denote the number of UEs associated with cell b in generation g. The per-generation association is modeled as the solution of the PF objective

$$\max_{\{b_u^g\}_{u \in \mathcal{U}}} \sum_{u \in \mathcal{U}} \log\left(\frac{W_{b_u^g}}{K_{b_u^g}^g}\, \mu_{b_u^g, u}\right). \tag{16}$$

3) Solver of optimization: The optimization problems (5),(9), (12), and (16) are mixed-integer nonlinear programming problems. We use the open-source optimizers [29], [30] and select the solution with the best objective value for each trial. We run ten trials with different random seeds for each configuration. Assuming the SON-based parameter tuning [26], we select q_{offset} using the $N = 100$ case with a separate random seed from those used in the evaluation trials, and then fix $q_{\mathrm{offset}} = 6.6$ dB for all configurations.

D. Results

Fig. 4 shows the system utility L and system throughput for DC, ideal DS, FA-DS, ideal SG, and local SG. We define UE density as the number of UEs per BS, namely $N/(B^{\mathrm{L}} + B^{\mathrm{N}})$. As UE density increases, FA-DS degrades relative to DC and eventually becomes inferior to local SG. These results indicate

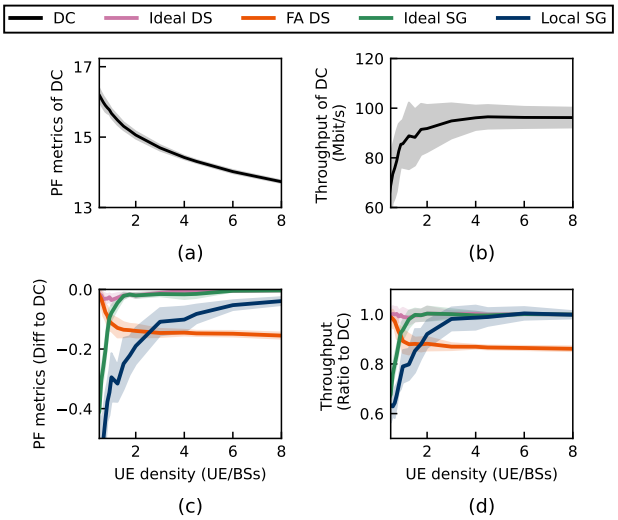

Fig. 4. System utility L and system throughput. Figs. (a) and (b) show the absolute results of DC. Figs. (c) and (d) show the performance of the ideal DS, FA-DS, ideal SG, and local SG relative to DC. The shaded area indicates the standard deviation over ten trials.

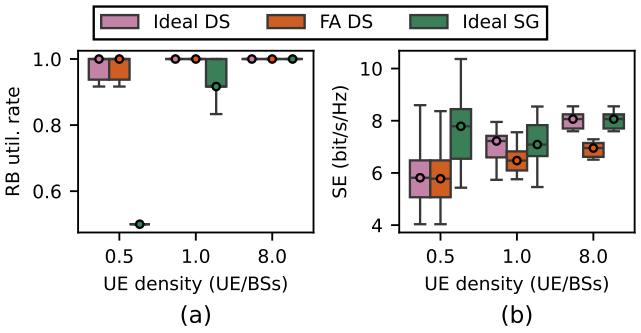

Fig. 5. System average RB utilization ratio and SE for each method.

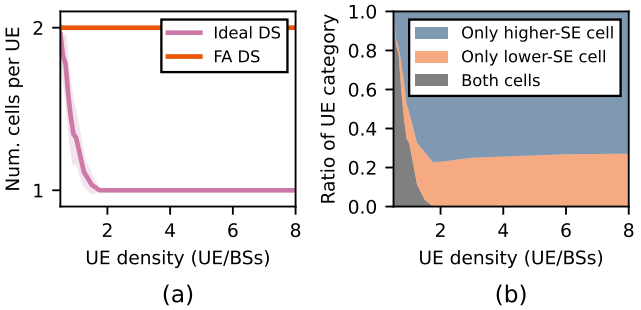

Fig. 6. UE-perspective spectrum analysis. (a) Number of cells from which the UE obtains bandwidth. (b) Ratio of UEs under ideal DS.

that persistent dual-path activation in DS can degrade system utility under high UE density, even at the flow level.

Figs. 4 (a) and (b) show the system utility L and system throughput of DC. Consistent with [14], the utility decreases, and the system throughput saturates as UE density increases. Figs. 4 (c) and (d) compare the other methods relative to DC. Ideal DS performs comparably to DC across UE densities. Both ideal SG and local SG approach DC as UE density increases, which is consistent with [14]. In the low-density regime, FA-DS is comparable to DC, whereas its relative performance degrades as UE density increases. In particular, FA-DS becomes inferior to local SG when UE density exceeds 3 UEs per BS.

The following explains the background that FA-DS performs well at low UE density but degrades at high UE density compared to the ideal DS and SG methods. Fig. 5 shows the system-average RB utilization ratio, and SE averaged over all BSs in both generations. As shown in Fig. 5 (a), ideal DS and FA-DS achieve RB utilization close to 1.0 at low UE density, which is higher than ideal SG. As UE density increases, the RB utilization ratio of SG also approaches 1.0. This occurs because DS activates up to two cells per UE across generations, leading to up to $2N$ active cell associations, whereas SG activates at most one cell per UE, leaving more cells unused in the low-density regime. This difference in the RB utilization ratio contributes to the high system utility of ideal and FA-DSs in the low UE density regime.

Fig. 5 (b) shows that, at high UE density, the system SE of FA-DS is lower than that of ideal DS and ideal SG. Since RB utilization saturates at 1.0 for all methods at high UE density, the lower SE of FA-DS directly leads to lower system utility.

To further analyze the SE degradation of FA-DS, Fig. 6 presents the UE-level perspective. Fig. 6 (a) shows the number of cells from which each UE obtains bandwidth. As UE density increases, ideal DS reduces the number of active cells per UE from two to one, whereas FA-DS keeps both cells active. Fig. 6 (b) shows the UE distribution under ideal DS by bandwidth source, namely the higher-SE cell only, the lower-SE cell only, or both candidate cells. For example, the lower-SE only means that the UE obtains bandwidth from a cell that is relatively lower-SE in the candidate cells, i.e., new and legacy cells. As UE density increases, ideal DS increasingly allocates a single higher-SE cell to each UE, improving system SE by balancing the load-SE trade-off. In contrast, FA-DS keeps both paths active for all UEs. As a result, each UE remains active on both a higher-SE cell and a lower-SE cell, and each cell shares resources equally among active UEs regardless of link quality. This forces the consumption of non-negligible resources on inefficient paths and decreases the system SE.

IV. Packet-level Simulation Evaluation on Existing Multipath Methods

To validate that the persistent splitting behavior assumed in FA-DS occurs in existing multipath transport schemes, we conduct a packet-level simulation. This section examines four multipath schedulers and three CCA methods in a simplified scenario and finds that they exhibit persistent splitting behavior, leading to inferior performance compared to the simple SG method. As discussed in Sec. III-B, the performance degradation of persistent splitting behavior at high UE density is owing to the inefficient bandwidth allocation between the

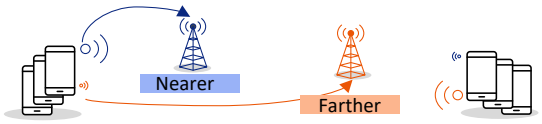

Fig. 7. System model for simple packet-level simulation. The six UEs are in the same position near the legacy BS and farther from the new BS. The other six UEs are at another position near the new BS.

UE's candidate paths with different SE. To clearly evaluate the occurrence and effect of persistent splitting in existing multipath mechanisms, this section emulates a situation in which persistent splitting causes significant performance degradation, i.e., the UE has two candidate paths with different SEs. Note that evaluation with a realistic setup is provided in Sec. V.

A. Setup

The simplified system consists of a legacy BS, a new BS, and 12 UEs, with the UE positions fixed, as illustrated in Fig. 7. Half of the UEs are located near the legacy BS, and the other half are located near the new BS. We conducted two scenarios. In scenario A, the SINR from the nearer and farther BSs is 15 dB and 5 dB, respectively. In scenario B, the SINR from the nearer and farther BSs is 30 dB and 0 dB, respectively. In each scenario, the application server sends full-buffered traffic using QUIC [31] with CUBIC CCA [15]. As a baseline, we use an SG method in which the UE only connects to the dominant BS.

We use an open-source packet-level simulator [6], [32]. The simulation parameters are as follows; The wired connection between UPF and BS has sufficient bandwidth and propagation delay of 10 ms, that between UPF and application server has sufficient bandwidth and propagation delay of 1 ms, the packet size is 1,500 byte, 1 ACK per 2 packets for both application QUIC protocol and MPQUIC subflow. We simulate 10 s. Further details on the packet-level simulator configuration follow [6].

1) Baseline multipath methods: This paper examines four existing multipath schedulers with uncoupled CCA with CUBIC [15] as follows:

- Minimum RTT (MinRTT) selects the path with the smallest RTT, which is the default scheduler of Linux MPTCP.
- Early completion first (ECF) [9] estimates the completion time of incoming data on each path and schedules packets on the path that minimizes the expected completion time.
- Blocking estimation (BLEST) scheduler [10] aims to avoid HoL blocking by preventing transmissions on slower paths when blocking is expected.
- Cross-layer information-based one-way delay predictive scheduler (CPS) [11] is a cross-layer scheduler that predicts the delay using the queue length and MAC level throughput, and selects the minimum delay path.

We examine two coupled CCAs with the MinRTT scheduler as follows.

- OLIA [12] is a coupled CCA designed for MPTCP, which dynamically adjusts the CWND of each subflow based on path conditions, increasing the window more aggressively on better paths and less on congested ones.
- Network-side multipath access management (NSMAM) [13] is a cross-layer coupled CCA, which defines the path priority, and suppresses the CWND of the non-priority path when the MAC queue of the WLAN AP on that path is occupied.

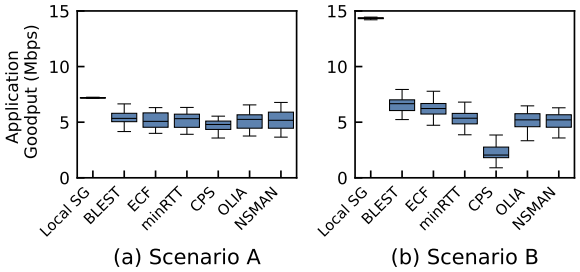

Fig. 8. Application level per-UE goodput.

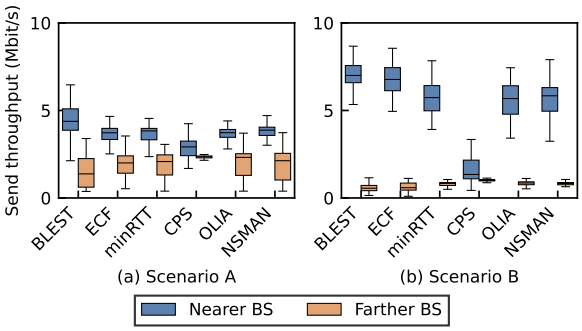

Fig. 9. Send throughput for paths with nearer and farther BSs.

TABLE I
BANDWIDTH ALLOCATION RATIO BETWEEN NEARER AND FARTHER BSS

	Scenario A		Scenario B	
Method	Nearer	Farther	Nearer	Farther
BLEST	59%	41%	45%	55%
ECF	49%	51%	43%	57%
minRTT	49%	51%	36%	64%
CPS	39%	61%	13%	87%
OLIA	49%	51%	36%	64%
NSMAM	51%	49%	36%	64%

2) Results: Fig. 8 shows the per-UE application-level goodput for each multipath method and local SG. Regardless of the scenarios, the SG method achieves higher goodput than any other multipath method. This is because, though the multipath method successfully suppresses the throughput of the farther cell, it remains in a splitting state. While a BS has packets in buffer to send to UEs, it allocates bandwidth equally among UEs to optimize intra-BS utilization. The BS consumes a large amount of bandwidth for non-dominant UEs, even though the throughput is low, resulting in reduced spectral efficiency. We support the analysis in the following paragraphs.

Fig. 9 shows the sending throughput of the multipath methods for the nearer and farther BSs per UE. The multipath scheduler successfully suppresses the traffic allocation to the

farther BS. Table I shows the ratio of the bandwidth that is allocated from nearer and farther BSs to UEs. Different from the sending throughput, large bandwidth is allocated from farther BS, at a minimum of 41%. This is because the BS with the PF scheduler allocates equal bandwidth to UEs with buffers. Even though the sending throughput is low, the buffer remains for a long duration due to the lower SE.

V. Proposed UE-density-aware DualSteer

To address the inefficiency of DS at high density, we propose a UE-DA DS that dynamically switches between single and splitting modes based on the number of active sessions on the secondary-path cell. The path priority is determined following the cell reselection criteria, which rely on the broadcast information in SIBs. UE-DA DS is self-contained within the UE and the UPF, since UE-DA DS relies only on information available at the UE, e.g., cell ID and cell-reselection offsets broadcast in SIBs, and on transport-layer observations at the UPF. Therefore, UE-DA DS preserves the fundamental DS requirement of no RAN impact. Moreover, the broadcast information used by UE-DA DS is updated at a coarse timescale, typically upon reselection and handover. UE-DA DS does not require frequent cross-layer signaling. The proposed UE-DA scheme is summarized in Fig. 10.

As a baseline, this paper uses the local SG method, in which the UE selects an appropriate generation according to the 3GPP-specified cell reselection procedure [25]. The criteria comprise UE-measured SINR and the cell offset broadcast in SIBs, including SIB1 and SIB4. This reselection criterion is also used in UE-DA DS to determine the priority and secondary paths.

A. Detail of proposed method

UE-DA DS defines two operation modes for each UE: single and splitting modes. In single mode, the UPF uses only the UE's priority path. In splitting mode, the UPF may use both the primary and secondary paths for the UE. This evaluation uses uncoupled CUBIC [15] and the minRTT scheduler for UE-DA DS. UE-DA DS is orthogonal to the underlying multipath scheduler and CCA.

The UE-DA DS consists of three phases: initial, exploring, and sustaining. In the initial phase, the UE and the UPF establish MPQUIC connectivity, determine a priority path based on cell reselection information, and begin communication on the priority path in single mode. During the exploration phase, the UPF estimates the UE density on the secondary-path cell by counting active sessions associated with that cell ID, and decides whether to enable split mode. When the mode changes, the sustaining phase keeps the selected mode for a sustain duration to prevent rapid oscillations; after the timer expires, UE-DA DS returns to the exploring phase. Note that the overhead of UE-DA DS is that the maintenace of the per-UE cell ID in UPF and UE notify its campon cell.

Initial phase: When UE starts the communication, the UE establishes two packet data unit (PDU) sessions via legacy and new BSs, and subsequently establishes two QUIC sessions

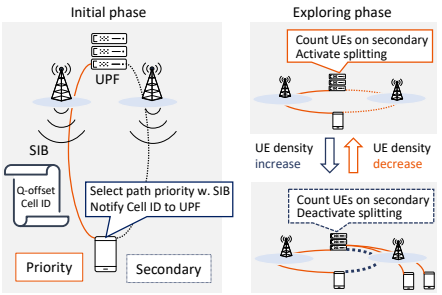

Fig. 10. Summary of proposed UE-DA DS.

with the MPQUIC proxy at the UPF [8], [19]. The legacy or new BS to which the UE connects is selected independently within each RAN according to its own selection mechanisms, e.g., cell reselection [25]. The UE determines the priority and secondary paths according to the cell reselection criteria, which is used when the UE becomes RRC active from idle [25]. Subsequently, UE sends the cell identifiers of the priority and secondary paths [25] to the UPF via application metadata on HTTP/3. Note that in 5G NR, NR cell global ID (NCGI) [33] can serve as an identifier, as it is a globally unique identifier for the cell and is broadcast by the BS to the UE via SIB1.

Exploring phase: During the exploring phase, the UPF periodically estimates the UE density of each cell on the secondary path, every Δ seconds, by counting the number of active sessions associated with that cell's ID. A session is regarded as active on a path if it has one or more ACK-eligible in-flight packets on that path. For a given UE whose secondary-path cell ID is b, let $k[b]$ denote the number of active sessions on cell b. UE-DA DS enables splitting if $k[b] < K$ and otherwise switches to single mode, where K is a load threshold. To avoid oscillation, once the mode changes, UE-DA DS keeps the new mode for a sustained duration before re-entering the exploring phase. The UPF maintains per-cell counters and per-UE mode states, and the update cost per decision epoch scales linearly with the number of UEs. Note that the UE notifies the cell id to the UPF, every time the camp-on cell is changed.

B. Simulation evaluation setup

We conduct a large-scale packet-level simulation with 12 BSs and up to 100 UEs. Unless otherwise specified, we reuse the physical and radio configurations, including UE and BS locations, bandwidth, propagation model, and local SG offset, from Sec. III-C, and the packet-level configuration, including packet size and wired-link delays, from Sec. IV. The serving cells of each generation are assumed to be optimized via each generation and obtained by Sec. III-C2.

We evaluate both static and dynamic scenarios. In the static scenario, UEs are fixed, and the application generates sufficient data to keep the downlink continuously backlogged. In the dynamic scenario, application of each UE starts communication gradually at a rate of 6.6 UEs/s, and each application stops communication after 15 s of activity. As a result, the UE density γ increases from $1/12$ UE/BS to $100/12$ UE/BS

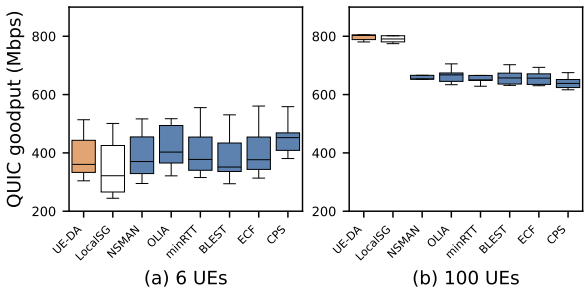

Fig. 11. System goodput

TABLE II
SYSTEM GOODPUT AS A FUNCTION OF UE DENSITY γ. **Boldface**, UNDERLINING, AND GRAY SHADING INDICATE THE BEST, SECOND-BEST, AND WORST VALUES, RESPECTIVELY.

Method	System goodput (Mbit/s) $\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 2.0$	$\gamma = 8.3$
UE-DA	391	**640**	**782**	**808**
Local SG	352	571	711	800
BLEST	389	592	686	660
CPS	**454**	614	688	641
ECF	410	563	685	657
minRTT	409	545	684	659
NSMAM	393	549	675	657
OLIA	420	560	663	665

and then decreases back to $1/12$ UE/BS within a trial. Each configuration is simulated for 30 s per trial over five trials with different random seed and diffrent UE locations.

For UE-DA, the sustain timer is randomly drawn from 100 to 500 ms, and Δ is set to 1 ms. Unless otherwise specified, we use $K = 1$ as a default, i.e., enabling splitting only when the secondary cell has no active session. A sensitivity study on K is provided at the bottom of Sec. V-C.

C. Results

Static scenario. Fig. 11 compares system goodput, i.e., the sum of per-UE QUIC goodput, for UE-DA DS, local SG, and DS with existing multipath methods, under several UE densities γ. At high UE density $\gamma = 8.3$, UE-DA achieves comparable performance to local SG and improves goodput by 21.5% over the best existing multipath baseline, i.e., OLIA. At low UE density ($\gamma = 0.5$), UE-DA achieves 11.2% higher throughput than local SG, while maintaining competitive goodput relative to existing multipath baselines. These results indicate that UE-DA retains bandwidth-aggregation gains at low UE density and avoids the performance loss at high UE density observed with conventional multipath DS.

Table II shows the system goodput as a function of the various UE densities γ. When γ is greater than 1, the proposed UE-DA achieves higher goodput than any other multipath method, comparable to local SG. At $\gamma = 0.5$, the local SG's throughput is lower than that of any other method, whereas the proposed UE-DA achieves performance comparable to that of other multipath methods. This is because, in low UE density, multipath bandwidth aggregation increases throughput by increasing RB utilization.

TABLE III
RB UTILIZATION RATIO AND SE

Method	$\gamma = 0.5$ RB util. rate	$\gamma = 0.5$ SE (bit/s/Hz)	$\gamma = 8.3$ RB util. rate	$\gamma = 8.3$ SE (bit/s/Hz)
UE-DA	57%	6.7	100%	7.8
Local SG	48%	7.0	100%	7.7
minRTT	60%	6.5	100%	6.6

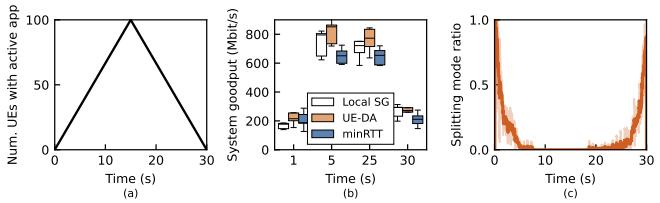

Fig. 12. Dynamic UE-density scenario evaluation. (a) Number of UE with active application over time. (b) System goodput at representative time instants (1 s, 5 s, 15 s, 25 s, and 30 s). (c) Splitting-mode ratio of UE-DA over time.

To explain these trends, Table III reports system RB utilization and SE. At $\gamma = 0.5$, multipath schemes, i.e., UE-DA and minRTT, exhibit higher RB utilization than local SG, which directly contributes to higher goodput in low-density regimes. In contrast, at $\gamma = 8.3$ the RB utilization saturates at nearly 100% for all methods, and performance becomes SE-limited. Here, UE-DA and local SG achieve higher SE than other multipath methods. This SE advantage stems from suppressing unnecessary dual-path activation on inefficient paths, i.e., avoiding excessive splitting, which we further quantify in the following dynamic and parameter studies.

Dynamic scenario. To evaluate the adaptability of UE-DA DS to time-varying UE density, we conduct a dynamic scenario in which UEs gradually start and stop communication, yielding an increase and decrease in UE density over time, as shown in Fig. 12 (a). Fig. 12 (b) reports system goodput at representative time instants, corresponding to low- to high-density time regimes. The average goodput over 30 s are 731 Mbit/s, 702 Mbit/s, and 570 Mbit/s for UE-DA DS, local SG, and DS with minRTT. In the early and later stages, i.e., 1 and 30 s, where UE density is lower than 2 UE/BS, the UE-DA and minRTT achieve higher goodput than local SG, benefiting from bandwidth aggregation. In the middle stage, i.e., 5 s to 25 s, UE density becomes high. DS with minRTT is inferior to the local SG due to persistent dual-path activity, whereas UE-DA suppresses excessive splitting and achieves performance comparable to local SG while outperforming minRTT. Fig. 12(c) further illustrates that the splitting-mode ratio of UE-DA. The splitting-mode ratio decreases in the early stages as the system approaches a high-density regime and increases in the latter stages as it approaches a low-density regime. Thus, we can confirm that UE-DA adapts the mode selection to UE-density dynamics and achieves the benefit reported in the static scenario.

TABLE IV
SYSTEM GOODPUT AND SPLITTING MODE RATIO FOR EACH K

Method	K	(System goodput (Mbit/s), Splitting mode ratio) $\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 8.3$
UE-DA	1	391, 100%	640, 49%	782,10%	808, 0%
	2	391, 100%	560, 100%	730, 52%	809, 0%
	3	391, 100%	554, 100%	689, 82%	809, 0%
Local SG	-	352, -	571,-	711,-	800, -
minRTT	-	409, -	545, -	684, -	659,-

D. Discussion on parameter K

Table IV summarizes the effect of K on UE-DA behavior. A larger K relaxes the condition for enabling splitting and thus tends to keep UE-DA in multipath operation, whereas a smaller K makes UE-DA more conservative and closer to local SG. UE-DA exhibits consistent behavior at $\gamma = 0.5$ and $\gamma = 8.3$ across all tested values of K. For example, at $\gamma = 0.5$, UE-DA consistently exploits bandwidth aggregation with 100% splitting and achieves identical goodput for all K, which is higher than local SG. The impact of K mainly appears at moderate densities. At $\gamma = 1$, $K = 1$ achieves the highest goodput with 49% splitting ratio, whereas larger K values keep splitting active all the time and reduce goodput, which is similar to the results of minRTT. Overall, in our simulation evaluation, $K = 1$ provides the most robust performance across the evaluated densities.

VI. CONCLUSION

This paper showed that DS can degrade system utility at high UE density due to persistent dual-path activation under RAN-unawareness by flow-level modeling. The packet-level simulations confirmed that existing L4 multipath mechanisms cannot avoid this inefficiency. To address this issue without impacting the RAN, we proposed UE-DA DS, which switches between single and splitting modes based on estimated cell load. Future work includes extending the evaluation to multi-band operation, UE mobility, and bursty traffic models, while keeping the method self-contained within the UE and the UPF.

REFERENCES

[1] C. Zhang, M. Dai, A. K. Salkintzis, D. Dimopoulos, H. Wang, and Y. Xu, "Migration matters: The shift from 5G to 6G," *J. ICT Standardization*, vol. 13, no. 3, pp. 281–300, 2025.

[2] GSMA, "5G-advanced: Shaping the future of operator services," GSMA Whitepaper, 2024, accessed: 2026-02-27. [Online]. Available: https://www.gsma.com/solutions-and-impact/technologies/networks/gsma_resources/

[3] 3GPP, "E-UTRA and NR; multi-connectivity; overall description; stage 2," *TS 37.340 ver. 18.5.0*, 2025.

[4] I. Mahmud, T. Lubna, and Y.-Z. Cho, "Performance evaluation of MPTCP on simultaneous use of 5G and 4G networks," *Sensors*, vol. 22, no. 19, p. 7509, 2022.

[5] T. Ogawara, S. Itahara, A. Suzuki, and M. Suzuki, "Toward sustainable 6G cellular system core-network-level traffic aggregation: An empirical study," *IEEE Access*, vol. 13, pp. 116 856–116 868, 2025.

[6] A. Suzuki, S. Itahara, T. Ogawara, and M. Suzuki, "Multi-generation Dualsteer: Cross-layer traffic control with asymmetric RAN feedback," *IEEE Access*, vol. 14, pp. 12 594–12 604, 2026.

[7] GSMA, "5G momentum continues with 1.6 billion connections worldwide, rising to 5.5 billion by 2030," GSMA White Paper, 2024, accessed: 2026-02-27. [Online]. Available: https://www.gsma.com/newsroom/press-release/

[8] 3GPP, "Procedures for the 5GS," *TS 23.502 ver. 18.9.0*, 2025.

[9] Y.-s. Lim, E. M. Nahum, D. Towsley, and R. J. Gibbens, "ECF: An MPTCP path scheduler to manage heterogeneous paths," in *Proc. CoNEXT*, Nov. 2017, pp. 147–159.

[10] S. Ferlin, Ö. Alay, O. Mehani, and R. Boreli, "BLEST: Blocking estimation-based MPTCP scheduler for heterogeneous networks," in *Proc. IFIP networking*, May 2016, pp. 431–439.

[11] B. Zhao, W. Yang, W. Du, Y. Ren, J. Sun, Q. Wu, and X. Zhou, "A multipath scheduler based on cross-layer information for low-delay applications in 5G edge networks," *Comput. Netw.*, vol. 244, p. 110333, 2024.

[12] R. Khalili, N. Gast, M. Popovic, and J.-Y. Le Boudec, "MPTCP is not Pareto-optimal: Performance issues and a possible solution," *IEEE/ACM Trans. Netw.*, vol. 21, no. 5, pp. 1651–1665, 2013.

[13] K. Chen, X. Xing, M. R. Palash, J. Liu, and J. Martin, "Network-side multipath access management in wireless networks with software-defined networking," *IEEE Trans. Veh. Tech.*, vol. 68, no. 10, pp. 10 030–10 044, 2019.

[14] A. Ghasemi, N. Limam, R. Boutaba, and A. Saleh, "Multi-connectivity for enhanced throughput: a critical study," in *Proc. IEEE NOMS*, May 2025, pp. 1–9.

[15] L. Xu, S. Ha, I. Rhee, V. Goel, and L. Eggert, "CUBIC for Fast and Long-Distance Networks," RFC 9438, 2023.

[16] H. Wang, C. Rosa, and K. I. Pedersen, "Inter-eNB flow control for heterogeneous networks with dual connectivity," in *Proc. IEEE VTC Spring*, May 2015, pp. 1–5.

[17] L. Weedage, C. Stegehuis, and S. Bayhan, "Impact of multi-connectivity on channel capacity and outage probability in wireless networks," *IEEE Trans. Veh. Tech.*, vol. 72, no. 6, pp. 7973–7986, 2023.

[18] 3GPP, "System architecture for the 5GS," *TS 23.501 ver. 18.9.0*, 2025.

[19] Y. Liu, Y. Ma, Q. D. Coninck, O. Bonaventure, C. Huitema, and M. Kühlewind, "Multipath extension for QUIC," IETF, Internet-Draft draft-ietf-quic-multipath-14, Apr. 2025.

[20] T. Pauly, D. Schinazi, A. Chernyakhovsky, M. Kühlewind, and M. Westerlund, "Proxying IP in HTTP," RFC 9484, 2023.

[21] A. Ford, C. Raiciu, M. J. Handley, O. Bonaventure, and C. Paasch, "TCP Extensions for Multipath Operation with Multiple Addresses," RFC 8684, 2020.

[22] M. Gapeyenko, V. Petrov, D. Moltchanov, M. R. Akdeniz, S. Andreev, N. Himayat, and Y. Koucheryavy, "On the degree of multi-connectivity in 5G millimeter-wave cellular urban deployments," *IEEE Trans. Veh. Tech.*, vol. 68, no. 2, pp. 1973–1978, 2018.

[23] ITU, "Framework and overall objectives of the future development of IMT for 2030 and beyond." *Recommendation M.2160.*, 2023.

[24] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surv. & Tutor.*, vol. 15, no. 2, pp. 678–700, 2012.

[25] 3GPP, "E-UTRAN; UE procedures in idle mode," *TS 36.304 ver. 16.2.0*, 2020.

[26] O. Østerbø and O. Grøndalen, "Benefits of self-organizing networks (SON) for mobile operators," *J. Comput. Netw. and Commun.*, vol. 2012, no. 1, p. 862527, 2012.

[27] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," *TS 38.901 ver. 19.1.0*, 2025.

[28] M. U. A. Siddiqui, F. Qamar, F. Ahmed, Q. N. Nguyen, and R. Hassan, "Interference management in 5G and beyond network: Requirements, challenges and future directions," *IEEE Access*, vol. 9, pp. 68 932–68 965, 2021.

[29] P. Belotti, J. Lee, L. Liberti, F. Margot, and A. Wächter, "Branching and bounds tightening techniques for non-convex minlp," *Optim. Methods & Softw.*, vol. 24, no. 4-5, pp. 597–634, 2009.

[30] P. Bonami, L. T. Biegler, A. R. Conn, G. Cornuéjols, I. E. Grossmann, C. D. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya *et al.*, "An algorithmic framework for convex mixed integer nonlinear programs," *Discrete optim.*, vol. 5, no. 2, pp. 186–204, 2008.

[31] J. Iyengar and M. Thomson, "QUIC: A UDP-Based Multiplexed and Secure Transport," RFC 9000, 2021.

[32] "DSSIM: DualSteer Simulator," https://github.com/kr_mcg/DSSIM, 2025, accessed: 2026-02-21.

[33] 3GPP, "NR and NG-RAN Overall description; Stage-2," *TS 38.300 ver. 18.5.0*, 2025.