

Joint U-Plane and C-Plane Evaluation of Distributed UPF Deployment: Trade-offs and Practical Guidelines for 5G and Beyond MEC

1st Takumi ISHIHARA
Keio University
Kanagawa, Japan
takuan@sfc.wide.ad.jp

2nd Tomoaki KANAYA
SoftBank Corp
Tokyo, Japan
tomoaki.kanaya@g.softbank.co.jp

3rd Noriaki YAMADA
SoftBank Corp
Tokyo, Japan
noriaki.yamada@g.softbank.co.jp

4th Takashi NAKAMURA
SoftBank Corp
Tokyo, Japan
takashi.nakamura@g.softbank.co.jp

5th Keiichi SHIMA
SoftBank Corp
Tokyo, Japan
keiichi.shima@g.softbank.co.jp

6th Keisuke UEHARA
Keio University
Kanagawa, Japan
kei@wide.ad.jp

Abstract—Distributed User Plane Functions (UPFs) are essential for low-latency 5G Multi-access Edge Computing (MEC), yet their user-plane (U-Plane) benefits and control-plane (C-Plane) costs have not been jointly quantified. This paper evaluates U-Plane latency and C-Plane signaling cost across five UPF deployment scenarios. The U-Plane is modeled on a hierarchical topology with 126 aggregation sites and validated against ns-3 with a 0.6% average RTT difference, while the C-Plane is evaluated using 3GPP-compliant signaling procedures for multiple MEC connectivity mechanisms. Results show diminishing U-Plane gains, whereas C-Plane signaling load grows approximately linearly with UPF and UE counts under the proportional crossing-rate assumption. An 8-site regional deployment achieves 98% satisfaction of the ≤ 20 ms latency target at only 17% of the C-Plane cost of prefecture-level deployment. Finally, we derive practical deployment guidelines for selecting UPF configurations and discuss scalability implications toward 6G.

Index Terms—5G core, distributed UPF, MEC, control plane scalability, deployment guidelines

I. INTRODUCTION

The integration of Multi-access Edge Computing (MEC) into 5G and future mobile networks has transformed the role of mobile core infrastructure. Where prior generations served primarily as transparent transport between the user equipment (UE) and the Internet, today's mobile core must also expose compute resources located inside the operator network [4], [7]. This shift creates a structural mismatch with the existing architecture: to access in-network compute resources, UE traffic must terminate at a User Plane Function (UPF) and be steered to an MEC host, yet UPFs in current commercial deployments are typically installed at only a handful of centralized locations per operator. The resulting path detours undermine the very latency benefits MEC is intended to provide.

Distributing UPFs closer to users and applications is therefore widely seen as essential for delivering low-latency MEC services in 5G and beyond. Prior work on UPF placement has demonstrated user-plane (U-Plane) gains under finer-grained

distribution, including reductions in end-to-end latency, alleviation of traffic concentration in the backhaul, and improved support for emerging applications such as vehicular safety, factory automation, and real-time analytics [8]–[10], [12] consistently.

However, the overall system impact of distributed UPFs cannot be assessed by examining the U-Plane alone. The implications for the 5G core control plane (C-Plane) remain insufficiently characterized. Existing 5G core C-Plane benchmarking studies have measured per-function load in isolation under varying session-establishment rates [13]–[15], but they do not model how C-Plane load varies as a function of UPF deployment topology. To the best of our knowledge, no prior work has quantified U-Plane gains and C-Plane costs jointly, under a unified framework, as complementary functions of UPF deployment granularity.

To support flexible access to distributed MEC resources, 3GPP defines three principal connectivity mechanisms in TS 23.548 [3]: Distributed Anchor Point (DA), Uplink Classifier (ULCL), and Local Area Data Network (LADN). These mechanisms differ in their handling of session continuity and IP-address transparency, but share a common foundation in the PDU session-based control model defined by TS 23.501/502 [1], [2]. As UPF deployments scale and become more dynamic, this shared abstraction may introduce scaling pressure on the C-Plane that no single mechanism choice can resolve.

In this paper, we investigate the joint behavior of U-Plane performance and C-Plane scalability as a function of UPF deployment granularity, with a view toward operator deployment in 5G and toward the design constraints that will shape future mobile core architectures. Rather than comparing individual mechanisms in isolation, we focus on identifying the deployment configurations that balance latency requirements against signaling cost, and on understanding how this

balance shifts as networks evolve.

Our contributions are as follows:

- 1) We define five UPF deployment scenarios (P0–P4) spanning fully centralized to highly distributed configurations, and evaluate U-Plane latency, backhaul traffic, and C-Plane signaling load under a unified analytical and simulation framework on a hierarchical operator topology with 126 aggregation sites.
- 2) We demonstrate a pronounced asymmetry between U-Plane gains and C-Plane costs: U-Plane improvements exhibit clear diminishing returns beyond a moderate level of distribution, while C-Plane signaling load grows linearly with both UPF count and UE count, irrespective of the connectivity mechanism used.
- 3) We provide practical deployment guidelines that map latency requirements to recommended UPF configurations and connectivity mechanisms, identifying a regional 8-site deployment as the practical sweet spot for general 5G MEC services, and discussing the architectural implications for finer-granularity deployment in 6G and beyond.

The remainder of the paper is organized as follows. Section II reviews the 3GPP MEC connectivity framework. Section III surveys related work. Sections IV and V present the C-Plane and U-Plane evaluations. Section VI integrates both dimensions and derives deployment guidelines. Section VII concludes.

II. BACKGROUND AND 3GPP MEC CONNECTIVITY

A. MEC Connectivity Mechanisms

To support flexible UE access to distributed MEC resources, 3GPP TS 23.548 [3] defines three principal connectivity mechanisms.

Distributed Anchor Point (DA) assigns a PSA-UPF per PDU session within a local site. When the UE moves and its serving area changes, the anchor UPF itself is relocated. The relocation procedure depends on the Session and Service Continuity (SSC) mode applied (Section II-B). IP address changes may occur, requiring applications to handle session re-establishment.

Uplink Classifier (ULCL) maintains a PDU session anchored at a central UPF while inserting an intermediate UPF (I-UPF) for local MEC traffic steering. The central PSA-UPF remains the IP anchor, and MEC-directed uplink traffic is selectively forwarded to the local I-UPF. No UE reconfiguration is required, making ULCL straightforward to deploy for existing UEs.

Local Area Data Network (LADN) / Multiple PDU Sessions allow the UE to maintain concurrent PDU sessions, with MEC-directed traffic carried over a dedicated session. From the MEC access perspective, this behaves equivalently to DA. When combined with LADN, access can be restricted to UEs within a specific geographic area.

Table I summarizes qualitative characteristics of the three mechanisms. Quantitative signaling-load and latency comparisons are deferred to Section IV.

TABLE I
QUALITATIVE COMPARISON OF MEC CONNECTIVITY MECHANISMS

Mechanism	UE Reconfig.	IP Change	Primary Use Case
DA (SSC Mode 2)	Yes	Yes	Wide-area services
DA (SSC Mode 3)	No	Temporary	Wide-area services
ULCL	No	No	Large-scale, legacy UE
LADN / Multi-PDU	Yes	Per session	Enterprise / area-specific

B. Session and Service Continuity Modes

DA and Multiple PDU Sessions support three SSC modes that differ in how session continuity is maintained during UE mobility. **SSC Mode 1** preserves the same PSA-UPF anchor throughout the session lifetime; the MEC anchor cannot be updated, making it unsuitable for mobility-aware MEC scenarios and outside the scope of this work. **SSC Mode 2** releases the existing PDU session and establishes a new one upon UE movement, allowing the network to select the optimal MEC anchor for the new location at the cost of a brief connectivity gap. **SSC Mode 3** establishes the new session before releasing the old one; combined with multi-homing, both sessions coexist temporarily, reducing the connectivity gap at the cost of increased resource consumption during overlap. This work evaluates SSC Mode 2 and Mode 3 as the modes relevant to mobility-driven C-Plane load.

C. Scope: Orthogonality to Radio-Layer Handover

The C-Plane events analyzed in this work are *serving-area boundary crossings*, which trigger PDU session anchor relocation in the mobile core. These are conceptually distinct from radio-layer handover events (Xn or N2 handovers) that maintain RRC connectivity within the radio access network. Radio-layer handovers occur on a much faster timescale and do not necessarily involve UPF relocation; they are extensively studied in the RAN-mobility literature [2] and are orthogonal to the core-network signaling load analyzed here.

For tractability, we abstract serving-area boundary crossings as a Poisson process. This abstraction is justified by three considerations. First, MEC use cases such as vehicular safety or in-factory robotics are inherently area-confined, making frequent inter-site crossings unlikely during active MEC service consumption. Second, UPF relocation during a crossing affects C-Plane signaling but not the total U-Plane traffic volume between UE and MEC, since the aggregate data exchanged is independent of relocation events. Third, the Poisson abstraction allows the C-Plane model to capture SMF load growth as a function of UPF count and mobility rate without requiring detailed physical-layer handover modeling. Detailed crossing-rate modeling under realistic mobility patterns is left for future work.

III. RELATED WORK

Research relevant to this paper spans three directions: UPF placement optimization, 5G core control-plane scalability, and MEC connectivity mechanism evaluation.

UPF placement and U-Plane optimization. The ETSI MEC framework established the basis for co-locating MEC applications with distributed UPFs [4], [5]. Prior studies have investigated joint UPF and edge-server placement for latency reduction [8], energy-efficient UPF placement [9], dynamic UPF reconfiguration and predictive session remapping [10], [11], and MEC data offloading under UPF sinking [12]. These studies primarily optimize U-Plane objectives, while C-Plane signaling cost as a function of UPF deployment granularity remains largely unexplored.

5G core control-plane scalability. Existing studies benchmark open-source 5G cores under increasing or bursty session-establishment workloads and identify the SMF as an important scalability bottleneck [13]–[15]. However, they evaluate C-Plane load independently of UPF deployment topology and do not quantify its relationship with distribution granularity.

MEC connectivity mechanism evaluation. Prior studies and standards describe ULCL, SSC modes, and LADN for 5G MEC integration [6], [16], primarily focusing on session continuity, mobility, and traffic steering rather than their relative C-Plane cost at deployment scale.

Positioning of this work. To the best of our knowledge, no prior study has jointly quantified U-Plane gains and C-Plane costs as functions of UPF deployment granularity under a unified framework. This work addresses this gap and derives deployment guidelines that map latency requirements to UPF configurations and MEC connectivity mechanisms.

IV. C-PLANE EVALUATION

A. Signaling Model

We analyze the C-Plane signaling procedures defined in 3GPP TS 23.502 [2] for three MEC connectivity mechanisms: Distributed Anchor Point (DA) with SSC Mode 2 and Mode 3, Uplink Classifier (ULCL), and Local Area Data Network (LADN). The signaling events of interest are triggered when a UE crosses a serving-area boundary, which represents the primary driver of C-Plane load in distributed environments.

For each mechanism, we count the total number of signaling messages exchanged per crossing event and the subset that reaches the Session Management Function (SMF), since the SMF is the central bottleneck in session-based control [13], [14]. The Procedure Completion Time (PCT)—defined as the elapsed time from the triggering event until the UE can access the MEC service—is also derived for each mechanism, based on fixed processing and propagation delays across UE, RAN, AMF, SMF, and UPF. Table II summarizes the per-crossing signaling characteristics.

The SMF signaling load per simulation interval is modeled as follows. We first define the following notation. N_{UPF} and N_{UE} denote the number of UPFs and UEs, respectively. M_{mob} is the per-UE area-crossing rate, and $P_{\text{reloc}}(N_{\text{UPF}})$ is the conditional probability that a crossing event triggers anchor relocation. Their product represents the effective per-UE relocation-event rate. The per-crossing SMF-reaching signal counts are S_{rel} and S_{est} for session release and establishment (used in DA Mode 2), S_{anchor} for new-anchor establishment

TABLE II
PER-CROSSING SIGNALING CHARACTERISTICS

Mechanism	Total Signals	PCT	UE Involved	SMF Signals
DA (SSC Mode 2)	8	55.5 ms	Yes	4
DA (SSC Mode 3)	6	33.0 ms	No	3
ULCL	6	33.0 ms	No	6
LADN	8+	$\geq 5,058$ ms*	Yes	5

*Includes 5 s timer expiry; signaling procedure itself is 33 ms.

TABLE III
C-PLANE SIMULATION PARAMETERS

Parameter	Value
Number of UEs	100, 500, 1,000
Number of UPFs	1, 2, 4, 8, 16, 32
Base mobility rate	1.0 crossings/min/UE
LADN timer	1, 5, 10, 30 s
Simulation duration	60 s per condition

(used in DA Mode 3), and S_{ulcl} per ULCL activation event. $E_{\text{event}}(N_{\text{UPF}}, \alpha)$ denotes the ULCL activation event rate as a function of N_{UPF} and MEC traffic ratio α . Using these terms, we express the total SMF signaling load L over the simulation interval for each mechanism as:

$$L_{\text{DA2}}(N_{\text{UPF}}) = M_{\text{mob}} \cdot P_{\text{reloc}}(N_{\text{UPF}}) \cdot (S_{\text{rel}} + S_{\text{est}}) \cdot N_{\text{UE}} \quad (1)$$

$$L_{\text{DA3}}(N_{\text{UPF}}) = M_{\text{mob}} \cdot P_{\text{reloc}}(N_{\text{UPF}}) \cdot S_{\text{anchor}} \cdot N_{\text{UE}} \quad (2)$$

$$L_{\text{ULCL}}(N_{\text{UPF}}, \alpha) = E_{\text{event}}(N_{\text{UPF}}, \alpha) \cdot S_{\text{ulcl}} \cdot N_{\text{UE}} \quad (3)$$

where L_{DA2} , L_{DA3} , and L_{ULCL} denote the loads for DA (SSC Mode 2), DA (SSC Mode 3), and ULCL, respectively. All three expressions share the structure $L \propto N_{\text{UE}} \cdot f(N_{\text{UPF}})$ where f is mechanism-dependent but linear in N_{UPF} under our crossing-rate assumption (Section IV-B).

B. Simulation Setup

We implement the C-Plane signaling chains in ns-3 [18] using a discrete-event simulation approach. In the simulation, the effective mobility-event rate represented by $M_{\text{mob}} P_{\text{reloc}}(N_{\text{UPF}})$ in the analytical model is instantiated as a Poisson process whose rate scales proportionally with the number of UPFs:

$$\lambda_{\text{cross}}(N_{\text{UPF}}) = M_{\text{base}} \cdot N_{\text{UPF}}. \quad (4)$$

where $M_{\text{base}} = 1.0$ crossings/min/UE. This proportionality represents a conservative upper bound, as actual crossing rates decrease with larger serving areas (fewer UPFs); the model provides a tractable worst-case baseline for scalability analysis. Key simulation parameters are summarized in Table III.

C. SMF Load vs. Number of UPFs

Fig. 1 shows the total SMF signaling load as the number of UPFs increases from 1 to 32 (with $N_{\text{UE}} = 100$). All three mechanisms exhibit approximately linear growth under the proportional crossing-rate assumption of Eq. (4), as predicted

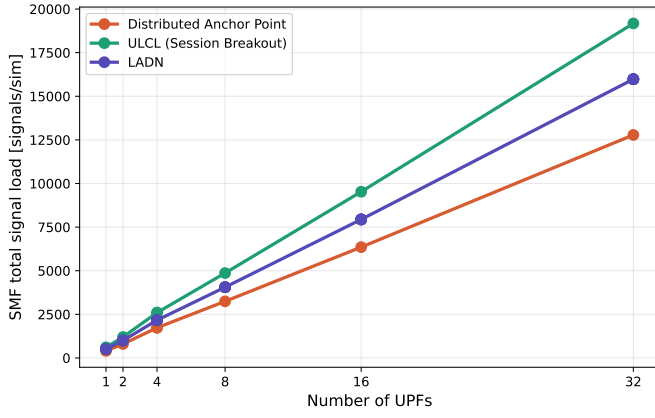


Fig. 1. SMF signaling load vs. number of UPFs ($N_{UE} = 100$). All three mechanisms scale approximately linearly with N_{UPF} under the proportional crossing-rate assumption; the slope is mechanism-dependent.

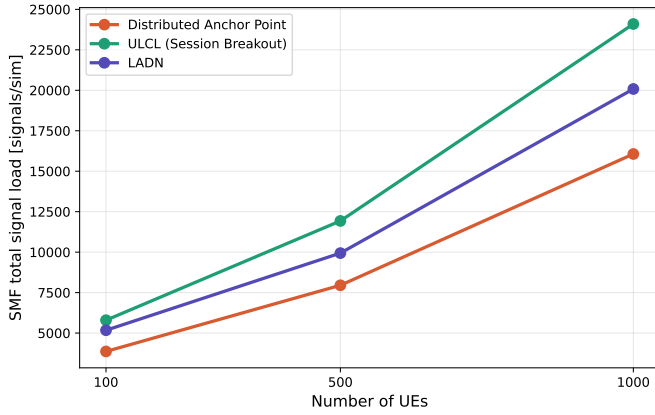


Fig. 2. SMF signaling load vs. number of UEs, averaged over all UPF configurations.

by Eqs. (1)–(3): as N_{UPF} increases, the effective mobility-event rate $\lambda_{cross}(N_{UPF})$ increases proportionally under Eq. (4), translating linearly into SMF load.

Among the three mechanisms, ULCL incurs the highest SMF load. Although ULCL requires only 6 signals per crossing, all I-UPF insertion and deletion events are routed through the SMF, giving it the highest per-crossing SMF-reaching signal count. DA (SSC Mode 2) sends 8 signals per crossing but only 4 reach the SMF, as the remaining 4 are handled between UE and RAN. LADN falls between DA and ULCL at large N_{UPF} due to its timer-driven session re-establishment behavior.

D. Scaling with UE Count

Fig. 2 shows SMF load as N_{UE} varies across 100, 500, and 1,000 (averaged over all UPF configurations). SMF load scales approximately linearly with UE count across all mechanisms, consistent with the N_{UE} factor in Eqs. (1)–(3).

Combined with the linear scaling in N_{UPF} established in Section IV-C, under the proportional crossing-rate assumption

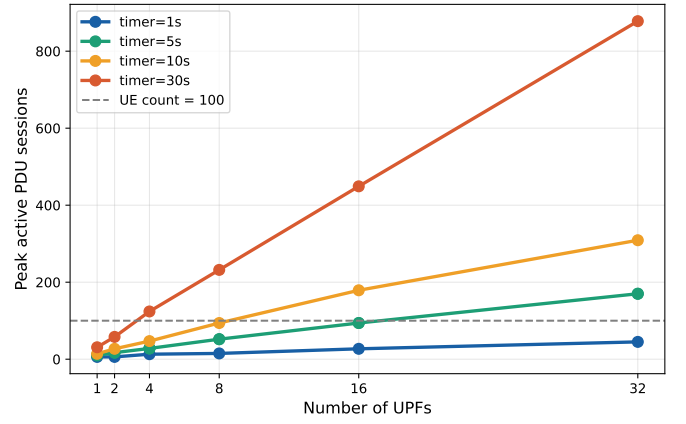


Fig. 3. LADN peak active PDU sessions vs. number of UPFs for varying timer lengths ($N_{UE} = 100$).

of Eq. (4), this reveals an approximately doubly-linear growth trend:

$$L_{SMF} \propto N_{UPF} \cdot N_{UE}. \quad (5)$$

A $10\times$ increase in both quantities yields approximately a $100\times$ increase in SMF load. Crucially, the same approximate scaling order is observed across all three connectivity mechanisms: mechanism selection adjusts only the proportionality constant, not the scaling exponent in N_{UPF} or N_{UE} . We return to the architectural implications of this observation in Section VI-D.

E. LADN Session Accumulation

Fig. 3 shows the peak active PDU session count for LADN under varying combinations of timer length and UPF count ($N_{UE} = 100$). With timer = 1 s, the peak count remains below 50 even at $N_{UPF} = 32$. With timer = 30 s, however, the peak reaches 878—approximately $8.78\times$ the UE count.

This session explosion occurs because UEs crossing area boundaries trigger new session establishments faster than the timer-based release mechanism can drain them, causing each UE to accumulate multiple simultaneous active sessions. The implication for deployment: at the P3 scale or above, LADN with any timer beyond a few seconds risks SMF session-state explosion.

F. Procedure Completion Time

PCT is independent of N_{UPF} for all mechanisms, as it is determined by the fixed signaling chain across UE, RAN, AMF, SMF, and UPF. DA (SSC Mode 2) yields $PCT = 55.5$ ms, while both ULCL and DA (SSC Mode 3) achieve 33.0 ms. ULCL achieves the same PCT as DA (SSC Mode 3) because it inserts an I-UPF without tearing down the existing IP anchor, whereas DA (SSC Mode 2) requires a two-phase session release and re-establishment. LADN incurs a PCT dominated by timer expiry: 5,058 ms at a 5 s timer, of which only 33 ms corresponds to actual signaling. Timer length thus directly controls the PCT-session-accumulation trade-off in LADN. These mechanism-level observations are integrated with U-Plane results in Section VI to derive deployment guidelines.

V. U-PLANE EVALUATION

A. Topology and Deployment Scenarios

We model a hierarchical mobile backhaul network as a rooted tree with five layers, comprising 126 aggregation sites and 181 links. The 126 sites arise from a five-layer expansion of 47 prefecture-level leaf nodes, with internal nodes representing higher-layer aggregation points. Each link is assigned a propagation delay estimated from inter-node distance assuming fiber-optic speed, and a bandwidth capacity assumed per hierarchy level; neither reflects any specific operator's measured parameters. The topology and deployment scenarios used here are representative of a country-scale operator network with multiple regional data centers and prefecture-level aggregation sites; the methodology is independent of the specific national topology and applies to any hierarchical operator network.

We define five UPF deployment scenarios (P0–P4) of increasing granularity, summarized in Table IV.

The number of CUs served per UPF decreases from 63 in P0/P1 to approximately 2.7 in P4, quantifying how finer granularity narrows the coverage area of each UPF instance. Since per-link propagation delay in the topology is assumed to scale with inter-node physical distance (Section V-A), this narrowing of coverage area is expected to correspond to shorter average UE–UPF physical distance as deployment granularity increases, although this relationship is not separately quantified in the present model.

P0 represents the current baseline in which MEC resources reside on the Internet side of the UPF, incurring full round-trip backhaul traversal. In P1–P4, MEC is co-located with UPF at progressively finer granularity. P0 and P1 share identical UPF placements, isolating the effect of MEC co-location from that of UPF distribution itself.

B. Analytical Model

Each CU generates MEC-directed traffic at rate

$$q_{\text{MEC}}(u) = \alpha \cdot n \cdot B \cdot \frac{8}{10^6} \quad [\text{Mbps}] \quad (6)$$

where $\alpha = 0.3$ is the MEC traffic ratio, $n = 100$ pkt/s is the per-CU packet generation rate, and $B = 1,500$ bytes is the packet size. Each CU is assigned to the UPF that minimizes accumulated propagation delay along the topology.

Link utilization is computed as

$$\rho(e) = \frac{\text{Traffic}(e)}{C(e)} \quad (7)$$

where $C(e)$ is the link capacity. Per-link queuing delay is modeled using an M/M/1 queue [17]:

$$d_{\text{queue}}(e) = \frac{\tau \cdot \rho(e)}{1 - \rho(e)} \quad (8)$$

where $\tau = B \cdot 8 / (C(e) \cdot 10^3)$ [ms/pkt] is the mean service time. The end-to-end RTT for CU u is

$$\text{RTT}(u) = 2 \cdot \left(d_{\text{wl}} + \sum_{e \in \text{path}(u)} (d_e + d_{\text{queue}}(e)) + d_{\text{mec}} \right) \quad (9)$$

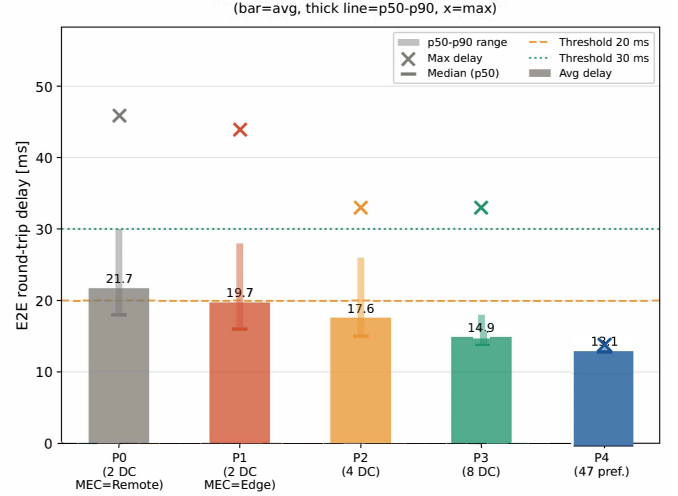


Fig. 4. E2E RTT distribution per deployment scenario (126-site topology, M/M/1 queuing model).

where $d_{\text{wl}} = 5$ ms is a fixed wireless delay and $d_{\text{mec}} = 1$ ms is the MEC processing delay.

Under the evaluated load conditions, the maximum link utilization is $\rho_{\text{max}} \approx 7 \times 10^{-5}$, placing all links firmly in the low-utilization regime where queuing delay is negligible; the dominant component of end-to-end latency is therefore propagation delay. We retain the M/M/1 formulation to validate (not assert) the low- ρ assumption and to remain applicable as α increases in future high-load scenarios.

C. Model Validation via ns-3

Packet-level simulation of the full 126-site topology in ns-3 is computationally prohibitive: at $n = 100$ pkt/s over a 60 s run, the total packet count exceeds 1.5×10^7 . We therefore select a representative subtopology of 12 aggregation sites across 8 administrative regions (a 1/10.5 scale reduction) and conduct packet-level simulation in ns-3. UDP flows are generated with Poisson arrivals ($\lambda = 100$ pkt/s per CU) and measured using FlowMonitor.

Table V compares the analytical model against ns-3. Average RTT agrees within 0.6% and maximum link utilization within 4.2%, validating the model's applicability for extrapolation to the full topology. The residual error stems from the difference between M/M/1 and the DropTail queue used in ns-3; at $\rho \approx 3 \times 10^{-4}$, queuing delay is essentially zero and the discrepancy is negligible.

D. End-to-End Latency Results

Fig. 4 shows the E2E RTT distribution across scenarios P0–P4, and Fig. 5 shows the corresponding latency-threshold satisfaction rates. Table VI summarizes key statistics.

The P0→P1 transition (2.0 ms improvement) demonstrates that co-locating MEC with the existing UPF—without changing the UPF placement itself—already provides measurable latency reduction. From P1 to P3, average latency improves by 4.8 ms as the number of regional UPF sites scales from 2

TABLE IV
UPF DEPLOYMENT SCENARIOS AND COVERAGE GRANULARITY

Scenario	UPF Placement	#UPFs	MEC Placement	Avg. CUs per UPF
P0	2 core DCs	2	Internet (centralized)	63.0
P1	2 core DCs	2	Co-located with UPF	63.0
P2	4 regional DCs	4	Co-located with UPF	31.5
P3	8 regional DCs	8	Co-located with UPF	15.8
P4	47 prefecture-level sites	47	Co-located with UPF	2.7

TABLE V
MODEL VALIDATION: ANALYTICAL VS. NS-3 (12-SITE SUBTOPOLOGY)

Metric	Model	ns-3	Error
Avg. E2E RTT	13.08 ms	13.00 ms	0.6%
Total traffic	~29 Mbps	~29 Mbps	~0%
Max link utilization	2.4×10^{-4}	2.5×10^{-4}	4.2%

TABLE VI
E2E LATENCY SUMMARY BY SCENARIO

Scenario	#UPFs	Avg	Max	≤ 20 ms	≤ 30 ms
P0	2	21.7 ms	46.0 ms	52%	90%
P1	2	19.7 ms	44.0 ms	52%	94%
P2	4	17.6 ms	33.0 ms	71%	98%
P3	8	14.9 ms	33.0 ms	98%	99%
P4	47	13.1 ms	14.0 ms	100%	100%

TABLE VII
TOTAL U-PLANE TRAFFIC AND REDUCTION PER SCENARIO

Scenario	Total [Mbps]	Vs. prev.	Vs. P0
P0	329.8	—	—
P1	239.0	−90.7 Mbps	−27%
P2	211.7	−27.4 Mbps	−36%
P3	181.4	−30.2 Mbps	−45%
P4	90.7	−90.7 Mbps	−72%

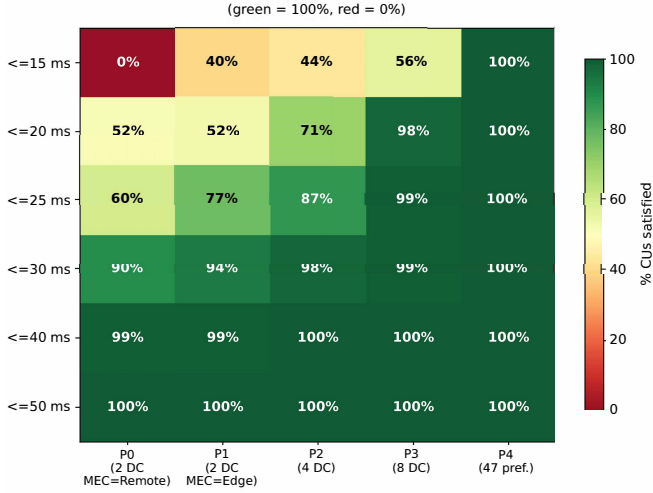


Fig. 5. Latency-threshold satisfaction rate per deployment scenario for representative thresholds ($\leq 15, 20, 25, 30, 40, 50$ ms).

to 8. P3 achieves 98% satisfaction of the ≤ 20 ms threshold, a widely adopted target for real-time MEC applications [7], while requiring only 8 UPF instances. Beyond P3, returns diminish sharply: P3→P4 adds 39 UPFs but yields only 1.8 ms of additional improvement, with ≤ 20 ms satisfaction rising by a mere 2 percentage points.

E. Backhaul Traffic Reduction

Total U-Plane traffic across all links decreases from 329.8 Mbps at P0 to 90.7 Mbps at P4, a reduction of approximately 72%. Table VII presents the per-scenario breakdown.

The P0→P1 transition alone accounts for 27% reduction by eliminating the full round-trip backhaul traversal to the Internet-side MEC host. This step requires no additional UPF hardware and incurs no C-Plane signaling cost increase, making it the highest-leverage first step in any MEC deployment roadmap. Operational costs associated with the MEC host co-location itself are outside the scope of this model.

F. Discussion

Two structural properties emerge. First, the majority of latency improvement is concentrated in the early deployment stages: the cumulative gain from P0 to P3 is 6.8 ms, whereas P3→P4 contributes only 1.8 ms despite a $6 \times$ increase in UPF count. This diminishing return arises because P3 resolves the primary geographic inefficiency—distant CUs no longer need to traverse multiple aggregation layers to reach a UPF. Second, the P0→P1 MEC co-location step provides disproportionate benefit relative to its deployment cost, suggesting it should be prioritized as the first step in any MEC deployment roadmap. These observations, combined with the C-Plane scaling structure of Section IV, motivate the integrated trade-off analysis of Section VI.

VI. INTEGRATED TRADE-OFF AND DESIGN GUIDELINES

Having quantified U-Plane and C-Plane behavior independently (Sections V and IV), we now integrate both dimensions to characterize the fundamental asymmetry of distributed UPF deployment and derive practical guidelines for operators.

A. Integrated Evaluation Framework

To compare deployment stages on a unified scale, we define the *C-Plane cost per millisecond of latency improvement*:

$$\kappa(P_i \rightarrow P_j) = \frac{\Delta L_{\text{SMF}}(P_i \rightarrow P_j)}{\Delta \text{RTT}(P_i \rightarrow P_j)} \quad [\text{signals/ms}] \quad (10)$$

where ΔL_{SMF} is the increase in SMF signaling load under the DA (SSC Mode 2) reference mechanism and ΔRTT is

the corresponding reduction in average end-to-end RTT. This metric expresses how expensive each additional millisecond of latency improvement becomes as distribution deepens, and allows a direct comparison between stages whose absolute values differ by orders of magnitude.

Table VIII summarizes both U-Plane and C-Plane indicators across all five scenarios. For P4 ($N_{\text{UPF}} = 47$), which lies beyond the simulated range of up to 32 UPFs, the C-Plane loads are analytically extrapolated using the linear scaling model under the proportional crossing-rate assumption. The structure of the asymmetry becomes apparent at a glance: $\overline{\text{RTT}}$ improves by less than 9ms across the entire P0–P4 range, whereas L_{SMF} grows by more than $23\times$.

B. Stage-by-Stage Analysis

P0→P1: zero-cost U-Plane improvement. The co-location of MEC with the existing UPF, without altering the UPF placement itself, incurs no additional C-Plane load ($\Delta L_{\text{SMF}} = 0$). Nevertheless, average RTT improves by 2.0ms and backhaul traffic decreases by 90.7Mbps (27%). This stage represents a pure architectural realignment with no mobility-driven signaling implications, and should be the first step in any MEC deployment roadmap.

P1→P3: cost-effective regional distribution. Expanding from 2 to 8 UPF sites yields a 4.8ms improvement in average RTT and raises ≤ 20 ms satisfaction from 52% to 98%. The per-stage cost is moderate:

$$\begin{aligned}\kappa(P_1 \rightarrow P_2) &= 808 / 2.1 \approx 385 \text{ signals/ms,} \\ \kappa(P_2 \rightarrow P_3) &= 1,616 / 2.7 \approx 599 \text{ signals/ms.}\end{aligned}$$

The C-Plane cost grows by a factor of ~ 1.6 between consecutive stages, while U-Plane gains remain substantial. The ≤ 20 ms requirement, which is widely adopted as a baseline target for vehicular and industrial MEC services [7], becomes nearly universally satisfied at P3.

P3→P4: steep asymmetry. Expanding from 8 to 47 sites adds 39 UPFs and increases SMF load by 15,756 signals ($5.9\times$ the P3 baseline), yet improves average RTT by only 1.8ms and ≤ 20 ms satisfaction by a mere 2 percentage points:

$$\kappa(P_3 \rightarrow P_4) = 15,756 / 1.8 \approx 8,753 \text{ signals/ms.}$$

This is approximately $23\times$ the cost-per-millisecond observed in the P1→P2 transition. The U-Plane is approaching the propagation-delay floor of the topology, while the C-Plane continues to scale linearly with UPF count—a regime where further distribution buys little latency at substantial signaling expense.

C. Deployment Guidelines

The regional sweet spot. Under the evaluated topology and traffic assumptions, P3 emerges as a practical sweet spot for general-purpose 5G MEC deployment. It achieves 98% ≤ 20 ms satisfaction at 17% of the C-Plane load of P4 (3,232 vs. 18,988 signals), while requiring only 8 UPF instances that can be co-located with existing regional data centers. Advancing from P3 to P4 raises the ≤ 20 ms satisfaction rate by only 2

percentage points, but raises the ≤ 15 ms satisfaction rate from 56% to 100%. Hence, P4 is justifiable only when applications with a hard ≤ 15 ms requirement—such as remote haptic control or tactile internet services [7] dominate the workload, and even then, it is contingent on substantial investment in SMF scalability.

Latency-requirement-driven selection. Table IX maps latency thresholds to recommended UPF configurations. The mapping is derived directly from the satisfaction-rate columns of Table VIII: the smallest UPF count that achieves at least 98% satisfaction for the target threshold is selected.

Mechanism selection within a fixed deployment. For a chosen UPF deployment, the connectivity mechanism further shapes the C-Plane profile. At P3, ULCL incurs $1.5\times$ the SMF load of DA (4,848 vs. 3,232 signals) while delivering the same procedure completion time ($\text{PCT} = 33$ ms). DA (SSC Mode 3) minimizes SMF-reaching signals (3 per crossing) and is preferable when scalability is the primary concern. ULCL is preferable when a transparent handover without UE reconfiguration is required, at the cost of a higher SMF load. LADN should be restricted to controlled, area-specific deployments due to its timer-driven session accumulation behavior, which makes it unsuitable for general distributed environments at the P3 scale or above.

D. Implications for 6G Core Design

The asymmetry observed across P0–P4 appears consistently across the evaluated topology, MEC traffic ratio, and connectivity mechanisms in this work. It suggests a structural tendency of the 3GPP control abstraction: every serving-area crossing must traverse the SMF because PDU session state is centrally arbitrated. Consequently, under the proportional crossing-rate assumption, SMF load exhibits approximately $\mathcal{O}(N_{\text{UPF}} \times N_{\text{UE}})$ growth, with the proportionality constant determined by the mechanism but the scaling order invariant across mechanism choices. The doubly-linear scaling holds across the entire range evaluated; however, its practical impact emerges only beyond the regional level. At P3 ($N_{\text{UPF}} = 8$, $N_{\text{UE}} = 10^3$), the absolute SMF load remains within the capacity envelope of current 5G core implementations [13]. Beyond P3, the scaling may become the binding constraint on further distribution, potentially well before the U-Plane reaches its propagation-delay floor.

The implication for 6G core design is twofold. First, the P0→P3 trajectory—MEC co-location followed by regional UPF distribution—is achievable within the current PDU session-centric framework and should be pursued as the near-term roadmap. Second, any deployment that aspires to prefecture-scale or finer granularity will require architectural mechanisms that decouple the control abstraction from per-session, per-UPF signaling. Promising directions include flow-state-level distribution among UPFs, gossip-based local arbitration without SMF involvement, and the unification of session and slice management at coarser granularity. These directions, however, require revisiting the session abstraction

TABLE VIII
INTEGRATED U-PLANE / C-PLANE COMPARISON ($N_{UE} = 100$)

Scenario	#UPFs	$\overline{RTT}$	≤ 20 ms	≤ 15 ms	L_{SMF}^{DA}	L_{SMF}^{ULCL}	Total traffic
P0	2	21.7 ms	52%	0%	808	1,212	329.8 Mbps
P1	2	19.7 ms	52%	40%	808	1,212	239.0 Mbps
P2	4	17.6 ms	71%	44%	1,616	2,424	211.7 Mbps
P3	8	14.9 ms	98%	56%	3,232	4,848	181.4 Mbps
P4	47	13.1 ms	100%	100%	18,988	28,482	90.7 Mbps

TABLE IX
RECOMMENDED UPF CONFIGURATION BY LATENCY REQUIREMENT

Requirement	Config.	L_{SMF}^{DA}	Note
≤ 50 ms	P0	808	Baseline sufficient
≤ 30 ms	P2	1,616	Most cost-efficient
≤ 20 ms	P3	3,232	Practical sweet spot
≤ 15 ms	P4	18,988	Requires SMF scaling

itself rather than optimizing within it, and we identify them as essential future work for 6G core design.

VII. CONCLUSION

This paper investigated the joint U-Plane and C-Plane behavior of distributed UPF deployment for 5G MEC and beyond. We evaluated five deployment scenarios (P0–P4) on a hierarchical operator topology with 126 aggregation sites, combining ns-3-based C-Plane signaling simulation with an M/M/1 queuing analytical U-Plane model validated against packet-level simulation within 0.6% error on average RTT.

Three findings emerge from this evaluation. First, U-Plane improvements exhibit pronounced diminishing returns: an 8-site regional deployment (P3) achieves 98% satisfaction of the ≤ 20 ms latency target, while scaling to 47 finer-granularity sites (P4) adds only 1.8 ms of further improvement at nearly $6\times$ the C-Plane cost. Second, under the proportional crossing-rate assumption, C-Plane signaling load exhibits approximately $\mathcal{O}(N_{UPF} \cdot N_{UE})$ growth across all three evaluated MEC connectivity mechanisms; mechanism selection adjusts the proportionality constant but not the scaling order. Third, co-locating MEC with the existing UPF (P0→P1) provides immediate U-Plane gains—2.0 ms latency reduction and 27% backhaul traffic decrease—at zero C-Plane cost, and should be the first step in any MEC deployment roadmap.

These findings translate into practical deployment guidance: a regional 8-site configuration is the sweet spot for general 5G MEC services under ≤ 20 ms requirements, and the choice of connectivity mechanism should be guided by the dominant constraint—SMF scalability favors DA (SSC Mode 3), while transparent handover favors ULCL. Looking toward 6G and beyond, the doubly-linear scaling of SMF load may become the binding constraint on further UPF distribution, potentially well before the U-Plane reaches its propagation-delay floor. Decoupling the control abstraction from per-session, per-UPF signaling emerges as an essential direction for future mobile core design.

REFERENCES

- [1] 3GPP, “System Architecture for the 5G System (5GS),” TS 23.501, Release 17, 2022.
- [2] 3GPP, “Procedures for the 5G System (5GS),” TS 23.502, Release 17, 2022.
- [3] 3GPP, “5G System Enhancements for Edge Computing,” TS 23.548, Release 17, 2022.
- [4] ETSI, “Mobile Edge Computing (MEC); Framework and Reference Architecture,” ETSI GS MEC 003, 2016.
- [5] ETSI, “Mobile Edge Computing (MEC); Mobile Edge Management; Part 2: Application lifecycle, rules and requirements management,” ETSI GS MEC 010-2, 2017.
- [6] ETSI, “MEC in 5G Networks,” ETSI White Paper No. 28, 2018.
- [7] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A Survey on Mobile Edge Computing: The Communication Perspective,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [8] Y. Li, X. Ma, M. Xu, and A. Zhou, “Joint Placement of UPF and Edge Server for 6G Network,” *IEEE Internet Things J.*, vol. 8, no. 22, pp. 16370–16378, 2021.
- [9] N. Das, K. Sarkar, T. Bose, A. Adhya, and S. Misra, “DeMPUP: Energy-Efficient UPF Placement for Beyond 5G Social Metaverse Traffic,” in *Proc. IEEE GLOBECOM*, Dec. 2024, pp. 4696–4701.
- [10] M. Brau et al., “Dynamic Scheduling and Optimal Reconfiguration of UPF Placement in 5G Networks,” in *Proc. ACM MSWiM*, 2020, pp. 103–111.
- [11] B. Gonzalez et al., “An Intelligent Scheduling for 5G User Plane Function Placement and Chaining Reconfiguration,” *Comput. Netw.*, vol. 235, 2023.
- [12] H. Jia et al., “MEC Data Offloading Strategy for UPF Sinking in 5G Core Network,” in *Lecture Notes in Electrical Engineering*, vol. 1186. Springer, 2024.
- [13] T. Mukute, E. Modroiu, and T. Magedanz, “Control Plane Performance Benchmarking and Feature Analysis of Popular Open-Source 5G Core Networks,” *IEEE Access*, 2024.
- [14] G. Lando et al., “SGC-Bench: A Framework for Stress-Testing and Benchmarking 5G Core VNFs,” arXiv:2509.18443, 2025.
- [15] T. O. Atalay, D. Stojadinovic, A. Stavrou, and H. Wang, “Scaling Network Slices with a 5G Testbed: A Resource Consumption Study,” in *Proc. IEEE WCNC*, 2022.
- [16] B. Jun et al., “Ultra-Low-Latency Services in 5G Systems: A Perspective from 3GPP Standards,” *ETRI J.*, 2020.
- [17] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. Wiley-Interscience, 1975.
- [18] ns-3 Consortium, “ns-3 Network Simulator,” <https://www.nsnam.org/>, accessed 2025.