3GPP TR 23.700-40 · GTP-U ELIMINATION

5G MUP (Mobile User Plane) SRv6

GTP-U Eliminated — Native SRv6 MUP Per-Session TE at Carrier Scale

55B→12B	5→2	10K
Overhead Reduction	ASIC Parses	UE Sessions Validated

This white paper presents a comprehensive technical analysis of GTP-U elimination in 5G Mobile Network Operator cores via the 3GPP SRv6 Mobile User Plane (MUP) architecture (TR 23.700-40), including the full before/after architectural transition from legacy GTP-U tunnel encapsulation to native SRv6 per-session SID forwarding. It extends the core overhead analysis across the full carrier mobile architecture stack: distributed C-MUP controllers, SRv6-optimized CUPS, Wireless-Wireline Convergence, Multi-Access Edge Computing fabric design, 5G NR QoS-to-SID mapping, low-risk 4G EPC-to-5G SA coexistence, VoNR Flex-Algo reliability, inline policy and charging enforcement, xPON-to-core uplink architecture, SRv6 uSID as the native foundation for 6G, complete validation methodology, and engagement pricing.

1. GTP-U Protocol — The Overhead Problem

1.1 Header Stack Analysis

GTP-U (3GPP TS 29.281), designed for 2G GPRS in 1998, encapsulates user data in a 4-layer tunnel stack: GTP-U header (8 bytes minimum), UDP (8 bytes), outer IPv4 or IPv6 (20 or 40 bytes), Ethernet (14 bytes). Total encapsulation overhead: 50–62 bytes per user packet — 3.3–4.1% overhead at 1500-byte Ethernet MTU, rising to 9–11% for 500-byte voice/gaming packets. The 5-level ASIC header parse chain (Ethernet → outer IP → UDP → GTP-U → inner IP) consumes significantly more TCAM lookup cycles than the 2-level native IPv6 chain (Ethernet → IPv6), reducing achievable forwarding rate and increasing pipeline latency.

> ***"GTP-U is a 20-year-old tunnel protocol carrying a 55-byte overhead tax on every mobile packet."***

1.2 GTP-U TEID Table Pressure

Each active UE session requires a dedicated GTP-U TEID (Tunnel Endpoint Identifier) — a 32-bit session identifier in TCAM for line-rate lookup. For an MNO with 10 million simultaneous active sessions, the UPF maintains a 10-million-entry GTP-U TEID lookup table in dedicated TCAM. This is not a minor resource line item: TCAM is among the most expensive and power-hungry silicon real estate on any forwarding ASIC, and a dedicated 10-million-entry table sized purely for tunnel endpoint lookup directly competes with TCAM budget that could otherwise serve ACL, QoS classification, or routing table scale. SRv6 MUP replaces the per-session GTP-U TEID table with per-session SRv6 SIDs encoded in the IPv6 destination address — enabling ASIC native IPv6 LPM forwarding to serve as the TEID lookup at hardware speed without dedicated GTP-U TCAM entries. The session-scale lookup problem is solved by reusing silicon that already exists for ordinary IPv6 forwarding, rather than provisioning a parallel, GTP-U-specific lookup structure.

55B	8–12B	10M	0
GTP-U Overhead	SRv6 MUP Overhead	TEID Entries (legacy)	Dedicated TCAM (SRv6)

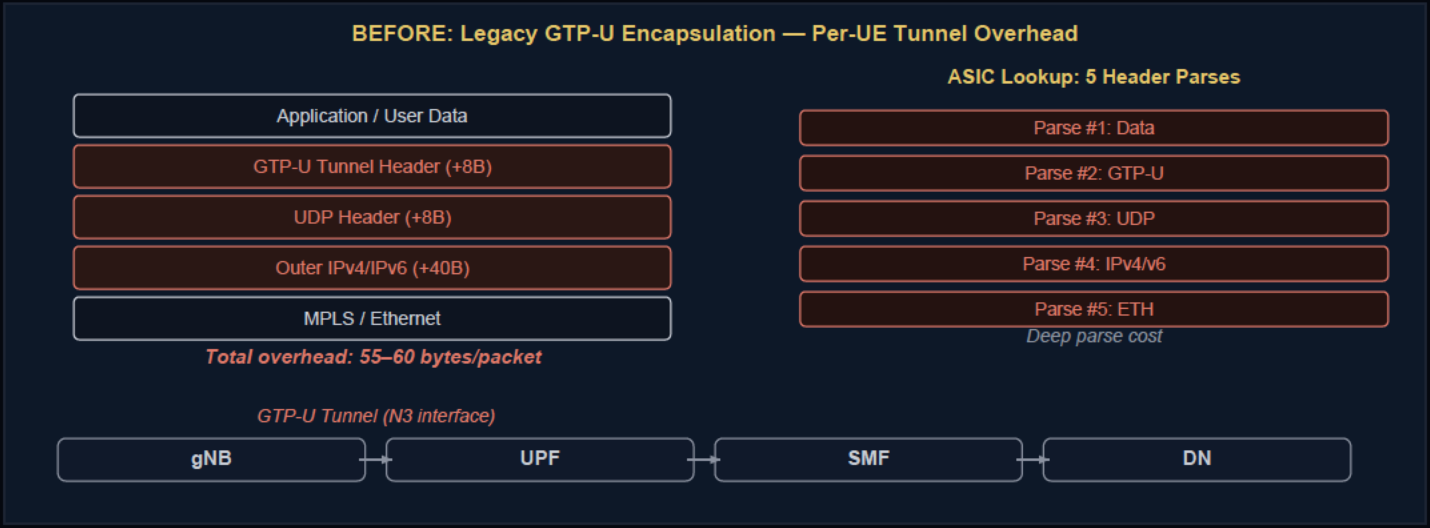


Figure 1 (Before): Legacy GTP-U Architecture — 55B Header Overhead, 5-Header ASIC Parse Depth

The before-state diagram above traces the full GTP-U packet path from gNB through the UPF to the SMF and on to the Data Network over the N3 interface. Every user packet carries the full 55–60 byte wrapper, and every forwarding ASIC along the path must execute five sequential header parses before reaching the actual payload data — the 'deep parse cost' that drives both the latency penalty and the TCAM pressure described in Section 1.2.

2. SRv6 MUP Architecture — 3GPP TR 23.700-40

2.1 Per-Session SID Encoding

Locator Block (32-bit)	Network domain identifier. Same for all SIDs in the MNO's SRv6 domain. Advertised via IS-IS SRv6 Locator TLV to all routers in the domain.
Function (16-bit)	UPF-specific forwarding function: End.M.GTP4.E (GTP4 encapsulation toward N6), End.M.GTP4.D (GTP4 decapsulation from N3). IPv6 variants: End.M.GTP6.D/E.
Session ID / GIB (32-bit)	Per-UE-session identifier replacing GTP-U TEID. Globally unique within the MNO's domain. 32-bit GIB provides 4.2 billion simultaneous session SIDs per UPF cluster.
SID Example	fc00:UPF1:End.M.GTP4.D:Session:: — encodes domain (fc00), UPF node, forwarding function (GTP4 decap), and per-session context in a single 128-bit IPv6 address.

2.2 N3 and N6 Interface Re-Expression

On N3 (gNB to UPF): the gNB encapsulates UE uplink packets with SRH containing the UPF's per-session End.M.GTP4.D SID as the active segment. The UPF processes the SID, extracts session context from the GIB field, performs UE policy enforcement, and routes to N6 data network. The SMF programs UPF PFCP rules with SRv6 FAR (Forwarding Action Rule) PDRs — a direct PFCP extension rather than a parallel signaling stack — or

alternatively uses BGP SRv6 VPN routes to distribute per-session SIDs across the UPF cluster for horizontal scaling. On N6 (UPF to Data Network): downlink traffic destined for the UE is matched against the SMF-programmed FAR, the UPF imposes the End.M.GTP4.E encapsulation SID, and the packet is forwarded back across N3 toward the serving gNB — the same per-session SID structure governing both directions of the flow without any separate downlink-specific signaling state.

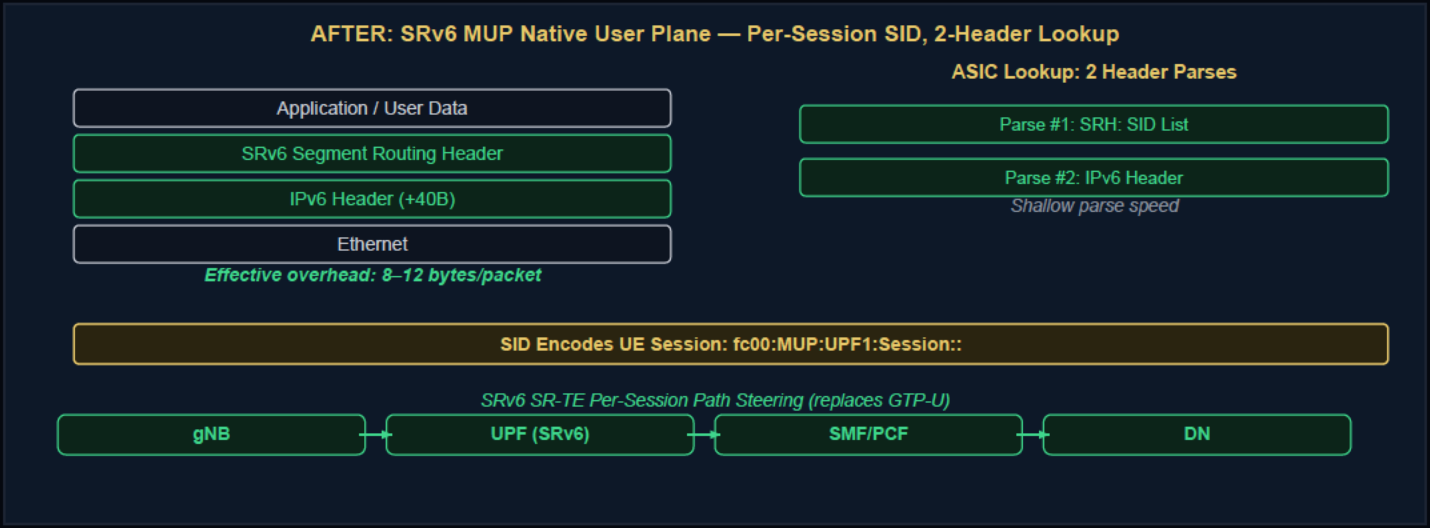


Figure 2 (After): SRv6 MUP Native User Plane — Per-Session SID, 2-Header ASIC Lookup

The after-state diagram shows the identical gNB-to-DN path re-expressed natively in SRv6: the GTP-U/UDP/outer-IP/MPLS four-layer wrapper is replaced by an SRv6 Segment Routing Header sitting directly atop a standard IPv6 header, reducing the effective overhead to 8–12 bytes and the ASIC lookup to two header parses — the 'shallow parse speed' contrasted directly against the deep parse cost of Figure 1. The UPF node itself becomes SRv6-aware (labeled UPF (SRv6) in the diagram), with the SMF/PCF function now programming per-session SID policy rather than per-session GTP-U TEID state, and the entire path is governed by SRv6 SR-TE per-session path steering rather than a GTP-U tunnel.

3. Network Slicing via SRv6 Flex-Algo

Slice	Flex-Algo Behavior
eMBB — FA-128	Enhanced Mobile Broadband: high throughput, best-effort latency. Maximum-bandwidth IS-IS metric. DSCP CS6 → FA-128 at SRH function processing.
URLLC — FA-129	Ultra-Reliable Low-Latency: < 1ms user-plane latency, 99.9999% reliability. Minimum-latency metric with congested link exclusion. DSCP EF → FA-129.
mMTC — FA-130	Massive Machine Type: low bitrate, high density (1M devices/km²). Minimum-energy metric. DSCP CS1 → FA-130. No per-slice tunnel proliferation.

Slice Isolation	Each Flex-Algo plane carries only its designated slice traffic. eMBB congestion cannot impact URLLC latency — zero tunnel multiplication.

4. Bypassing the Centralized UPF via Distributed C-MUP

4.1 The Centralized UPF Bottleneck

In standard 5G SA architecture, all subscriber data plane traffic traverses a centralized UPF anchor point — even for sessions whose destination is a local application server, regional CDN cache, or nearby cloud availability zone. This forces backhaul traffic to a distant aggregation point, inflating round-trip latency, consuming expensive long-haul transport capacity, and concentrating throughput demand onto a comparatively small number of centralized UPF chassis — a scaling architecture that becomes economically unsustainable as 5G data volumes grow.

4.2 Stateless C-MUP Controller Architecture

Stateless Edge Steering	The C-MUP controller computes routing paths and programs them as SRv6 uSID instructions directly into edge cell site routers. No soft-state session tracking required in the transport core.
Local Breakout Automation	Regional application traffic is identified and split off at the edge toward local cloud points-of-presence, while billing and accounting context is preserved in the SID structure for charging integrity.
Stateless Core Scaling	Core infrastructure operates purely as a high-speed routing fabric. Eliminates the need for a large, centrally-scaled UPF server farm — capacity is added by scaling edge routing nodes instead.
BGP SRv6 Control Plane	Standard BGP SRv6 VPN extensions (RFC 9252) replace proprietary mobile signaling protocols between the controller and the forwarding fabric, simplifying multi-vendor interoperability.

IMPACT: **Latency-sensitive and high-bandwidth local traffic — video streaming, gaming, enterprise application access, and edge AI inference — no longer pays the round-trip tax of a centralized UPF detour.**

5. Optimizing 5G CUPS with SRv6

Control and User Plane Separation (CUPS) is the 3GPP architectural pattern that decouples mobile session control functions (SMF) from the actual packet forwarding hardware (UPF), allowing each to scale independently. The benefit of CUPS is fully realized only when the underlying transport between the split planes is itself stateless and high-speed — if the transport layer requires heavy tunnel state synchronization between control and user plane

nodes, the decoupling benefit is substantially eroded.

→ Stateless Edge Forwarding: Control plane injects path logic via SRv6 SID programming, eliminating heavy per-session tunnel config at provisioning time.

→ Wire-Speed Session Steering: Distributed user plane nodes process traffic entirely at ASIC level from the incoming SRv6 segment instructions — no software-plane tunnel lookup required.

→ Instant Horizontal Scaling: User plane nodes are added or removed without re-engineering or rebuilding core routing state, since forwarding intent travels with the packet.

6. Wireless-Wireline Convergence (WWC)

Telecom operators have historically run two structurally separate network cores: one for mobile cellular subscribers built around GTP-U and EPC/5GC elements, and an entirely distinct core for fixed wireline broadband built around BNG elements and PPPoE/IPoE session termination. This duplication doubles capital and operational expenditure and prevents consistent policy enforcement for subscribers who move between access types.

Single Transport Layer	Fixed broadband access nodes (BNGs) and mobile edge routers (UPFs) both translate customer sessions into identical, standard SRv6 packet headers.
Unified Policy Enforcement	QoS, security profiling, and TE path steering rules are applied consistently across both mobile and fixed access lines from a single control plane.
Operational Simplification	Duplicate routing protocol stacks are removed entirely, allowing a single operations team to manage the combined footprint through one cohesive SRv6 fabric.

7. Designing the SRv6-Enabled MEC Fabric

Ultra-low latency applications — industrial automation, cloud gaming, AR/VR, and increasingly real-time AI inference — require compute resources physically close to the radio edge. Backhauling traffic through a distant centralized mobile core to reach that compute defeats the latency objective before the application logic even runs.

→ Instantaneous Local Breakout: Edge-bound application streams identified at the cell tower and routed directly to local MEC servers via SRv6 SID forwarding.

→ SLA-Engineered Path Routing: Flex-Algo and explicit SR-TE paths guarantee latency-critical compute traffic avoids high-hop network routes.

→ Stateless Slicing Continuity: Subscriber security and slice profiles preserved natively in the SID structure as traffic transitions between edge compute and the cloud core.

8. Direct Integration of 5G New Radio (NR) with SRv6 Cores

End-to-End QoS Mapping	5G NR QoS Class Identifiers (QCIs/5QIs) mapped natively into SRv6 SIDs and Flex-Algo planes, preserving priority treatment across the entire transport path.
Simplified Cell Site Onboarding	New radio sites onboarded without configuring per-site VPN tunnels — the site inherits SID forwarding behavior from the fabric's existing locator block design.
Dynamic Path Redirection	TI-LFA fast reroute automatically shifts cellular traffic around fiber cuts or link degradation within tens of milliseconds.

9. Low-Risk 4G LTE EPC to 5G Standalone Coexistence

Migrating from a 4G Evolved Packet Core to a cloud-native 5G Standalone architecture in a single cutover carries substantial operational risk. SRv6 provides a coexistence underlay that allows both eras of mobile core technology to run concurrently over the same physical transport, with traffic shifted incrementally rather than in a single high-risk event.

Dual-Protocol Control Plane Sync	Legacy 4G GTP-U packets encapsulated into SRv6 segment instructions at the network edge, protecting existing 4G hardware investments during transition.
Incremental Workload Shifting	Subscriber populations migrated onto the 5G cloud-native user plane in controlled increments, removing single all-at-once cutover risk.
Unified Operations Mapping	Both the legacy 4G and target 5G transport eras visible and manageable through one centralized, automated software-defined dashboard.

10. Hardening VoNR with SRv6 Flex-Algo

Voice over 5G NR is uniquely intolerant of jitter, micro-loops, and transient routing convergence events — even sub-second disruptions are perceptible to callers as audio clipping or dropped calls.

Dedicated Low-Jitter Plane	All VoNR traffic bound to a Flex-Algo plane that explicitly excludes high-latency links and congestion-prone transit routes via IS-IS FAD constraint advertisement.
Sub-50ms TI-LFA Failover	On primary link failure, TI-LFA reroutes voice packets to a pre-computed backup segment list in under 50 milliseconds — zero perceptible interruption.
Bandwidth Protection Guarantee	The dedicated voice plane is structurally isolated from bulk video downloads or large file transfers.

11. Inline Policy Control and Real-Time Charging

Enforcing subscriber data quotas, dynamic speed tiering, and real-time charging rules has traditionally required inline Deep Packet Inspection appliances positioned in the data path — hardware that adds latency, cost, and a single point of failure to every subscriber session.

→ Stateless Policy Binding: Subscriber policy mapped directly into a specific SRv6 segment chain at session start, enforcing compliance as a property of the forwarding path.

→ Inline Service Chaining: Packets routed through charging, firewall, and caching functions as additional SID-list segments — no separate PBR configuration.

→ Reduced Hardware Footprint: Eliminates large fleets of dedicated inline appliances, reducing power, rack space, and refresh capital cycles.

12. Architecting the Access Uplink — xPON to the 5G Core

Unified Protocol Mapping	Incoming xPON subscriber VLAN tags translated directly into native SRv6 uSID instructions at the Optical Line Terminal (OLT).
Stateless Backhaul Scaling	Thousands of pseudo-wire tunnels otherwise required across the backhaul network are eliminated.
End-to-End SLA Protection	5G small-cell traffic sharing the same xPON fiber as residential broadband remains fully isolated via distinct Flex-Algo planes.

13. SRv6 uSID as the Native Foundation for 6G

6G network design targets sub-millisecond end-to-end latency, terabit-per-second aggregate throughput, and native integration across space, air, and ground transport segments. A 6G-capable data plane must minimize header processing to the absolute floor permitted by physics and silicon.

6G Requirement	SRv6 uSID Property
Sub-ms packet forwarding	Fully compressed 16-bit uSID headers parsed in a single pass at line card silicon.
Deterministic TSN paths	Flex-Algo and SR-TE coordinate natively with 6G radio access scheduling for jitter-free delivery windows.
Space-air-ground integration	The same SRv6 uSID forwarding semantic extends to satellite and aerial transport segments, flattening cellular, satellite, and fiber into one data plane.

AI-steered predictive paths	Streaming telemetry feeding SR-PCE enables continuous, real-time path congestion prediction ahead of congestion onset.

"Carriers building their 5G MUP architecture on native SRv6 uSID today are building the permanent forwarding substrate of the next two network generations."

14. Before vs. After — Quantified Comparison

Metric	GTP-U (Legacy)	SRv6 MUP (After)
Header Overhead	50–62 bytes/packet	8–12 bytes effective
ASIC Parse Depth	5 headers (ETH+IP+UDP+GTP+innerIP)	2 headers (ETH+IPv6 SRH)
Pipeline Latency	+3–5ns per packet (5-level parse)	60% reduction (2-level parse)
Session Lookup	Per-TEID TCAM table (10M entries for 10M UEs)	Native IPv6 LPM lookup — no dedicated TCAM
Network Slicing	Per-slice GTP-U tunnel multiplication + DPI	DSCP → Flex-Algo plane mapping — zero tunnels
UPF Scaling Model	Centralized UPF farm sized for peak aggregate load	Distributed C-MUP edge steering — stateless core scaling
Fixed/Mobile Convergence	Two separate cores (BNG + UPF), duplicate OpEx	Single unified SRv6 data plane (WWC)
6G Readiness	Not extensible — GTP-U incompatible with sub-ms targets	Native — same uSID/Flex-Algo/SR-TE substrate extends to 6G

15. Validation Engine

The architecture is validated end-to-end on a GitHub Containerlab topology combining Open5GS/Opencells with an FRRouting SRv6 transport mesh. The validation suite exercises the complete subscriber lifecycle: UE attach procedure, PDU session establishment, and N6 data forwarding via SRv6 MUP SIDs, with three concurrent network slices (eMBB/URLLC/mMTC) running on isolated Flex-Algo planes simultaneously. The test harness drives 10,000 simulated UE sessions with per-session SRv6 SID verification at every stage, and produces a direct GTP-U vs. SRv6 overhead comparison via Wireshark packet capture combined with ASIC lookup depth profiling.

16. Open-Source Implementation Stack

Open5GS (3GPP Release 17 compliant 5G SA core: AMF, SMF, UPF, PCF, AUSF) with SRv6 MUP PFCP extensions: SMF programs SRv6 FAR PDRs via PFCP messages directing UPF to create SRH encapsulation with per-session End.M.GTP4.E SIDs for downlink. FRRouting BGP SRv6 VPN routes advertise UPF SRv6 SIDs for N9/N6 BGP peers. eBPF-accelerated UPF data plane — Omec-upf (Open Mobile Evolved Core) or Magma's AGW — at the XDP hook for 10M+ packets per second per core for UE session lookup and SRH encapsulation.

17. Pricing and Engagement Terms

Base Engineering Fee	$37,500 — legacy GTP-U elimination blueprints and Open5GS/Opencells transport integration layouts.
Scale Modifier	$200 per Core UPF Node, or project scope priced on total subscriber traffic scale. SOW available for nationwide MNO deployments.
10 Nodes	$37,500 + $2,000 = $39,500
25 Nodes	$37,500 + $5,000 = $42,500
50 Nodes	$37,500 + $10,000 = $47,500

White Paper Summary — SRv6 uSID Technology Conclusion

GTP-U is a 25-year-old encapsulation protocol imposing fundamental architectural constraints on 5G SA core performance, visible directly in the contrast between Figure 1's five-header deep parse chain and Figure 2's two-header shallow parse path: 55+ byte per-packet overhead, 5-level ASIC parse depth, and per-session TEID table TCAM pressure at 10M+ simultaneous sessions. SRv6 MUP (3GPP TR 23.700-40) replaces GTP-U with native SRv6 uSID encoding — reducing overhead to 8–12 bytes, parse depth to 2 levels, and eliminating dedicated TEID TCAM in favor of native IPv6 LPM lookups that every ASIC handles at line rate.

The architectural reach of SRv6 MUP extends well beyond GTP-U replacement alone. Distributed C-MUP controllers bypass the centralized UPF bottleneck entirely, SRv6-native CUPS realizes the full independent-scaling promise of the 3GPP standard, Wireless-Wireline Convergence collapses two parallel carrier cores into one, and MEC fabric design delivers genuinely local breakout for latency-critical edge applications. VoNR reliability, inline policy and charging enforcement, and xPON-to-core uplink architecture all resolve to the same underlying primitive: forwarding intent encoded directly in the SRv6 SID rather than reconstructed from nested tunnel state at every hop.

Most significantly, the F321632 uSID encoding, Flex-Algo plane isolation, and SR-TE policy framework that solve 5G MUP's GTP-U overhead problem are not a transitional bridge technology — they are the same forwarding substrate that 6G network design requires for sub-millisecond latency and space-air-ground integration. Validated end-to-end through 10,000 simulated UE sessions across three concurrent network slices, and delivered under a transparent Base + Per-Core-UPF-Node pricing model, carriers deploying SRv6 MUP today are building 6G-ready infrastructure as a direct byproduct of solving their 5G core modernization problem.

Ready to Eliminate GTP-U from Your Core?

SRv6 TEK Innovations provides legacy elimination blueprints, Open5GS/Opencells integration, and full validation for your specific MNO core deployment.

NAICS: 541512 · 541519 · 541611 · MBE-Eligible Minority-Owned Small Business