White Paper

Edge AI Workload Routing in Mobile Networks

March 2025

Research Institute
of Advanced Technology

SoftBank Corp.

SoftBank

Contents

1 Background

The rapid proliferation of AI services is driving an increasing demand for processing AI workloads at the mobile edge. Multi-access Edge Computing (MEC) offers a solution to handle these computationally intensive tasks closer to end-users, reducing end-to-end latency and load on the transport network, distribution of power consumption that mitigate a hurdle of building a massive datacenter by building smaller edge datacenters, and improving overall performance. AI-RAN [1] [2][3] is a new concept in MEC design that leverages high-performance common servers (e.g., GPU servers) to host both RAN software and AI software. Traditionally, these functions are deployed separately, requiring additional costs for common infrastructure. By using AI-RAN, we can run software-based RAN and AI workloads on the same distributed system by sharing common infrastructure, thereby enhancing the efficiency of investment, power consumption, and communication latency [4][5]. This shift towards edge AI processing is necessitated by the growing complexity of AI models, the need for real-time processing in applications like autonomous vehicles and augmented reality, and concerns over data privacy and sovereignty.

The importance of directing AI traffic to MEC lies in its potential to:

- Significantly reduce latency for time-sensitive AI applications
- Enhance data privacy and security by keeping sensitive information local
- Enable new classes of AI-driven applications that require edge processing

However, effectively routing AI traffic to MEC resources presents several challenges, including:

- Identifying and classifying AI traffic in real-time
- Ensuring service continuity as users move between different edge zones
- Integrating edge AI processing with existing network infrastructure

The 5G Core (5GC) network selects a User Plane Function (UPF) close to the User Equipment (UE) and forwards traffic to enable the local access to the Data Network (DN) (as described in Section 5.13 of TS23.501 [6]). There are three connectivity models to enable Edge Computing (as described in Section 4.3 of 3GPP TS23.548 [7]).

- Distributed Anchor Point: For a PDU session, the PDU Session Anchor UPF (PSA UPF) is in a Local Site, i.e. close to the UE location. This method does not have connectivity with Central Site.
- Session Breakout: A PDU session has a PSA UPF in a Central Site and one or more PSA UPF in the Local Site. The Central-PSA UPF provides the IP Anchor Point. The Edge Computing application traffic to Local Site is selectively diverted to the Local-PSA UPF using ULCL. The Local-PSA UPF may be changed due to e.g. UE mobility.
- Multiple PDU Sessions: Edge Computing applications use PDU session(s) with a PSA UPF(s) in Local Site. The rest of applications use PDU session(s) with PSA UPF(s) in Central Site. Any PSA UPF may be

changed due to e.g. UE mobility.

These three connectivity models are illustrated in Figure 1 (described in Section 4.3 of 3GPP TS 23.548 [7]).

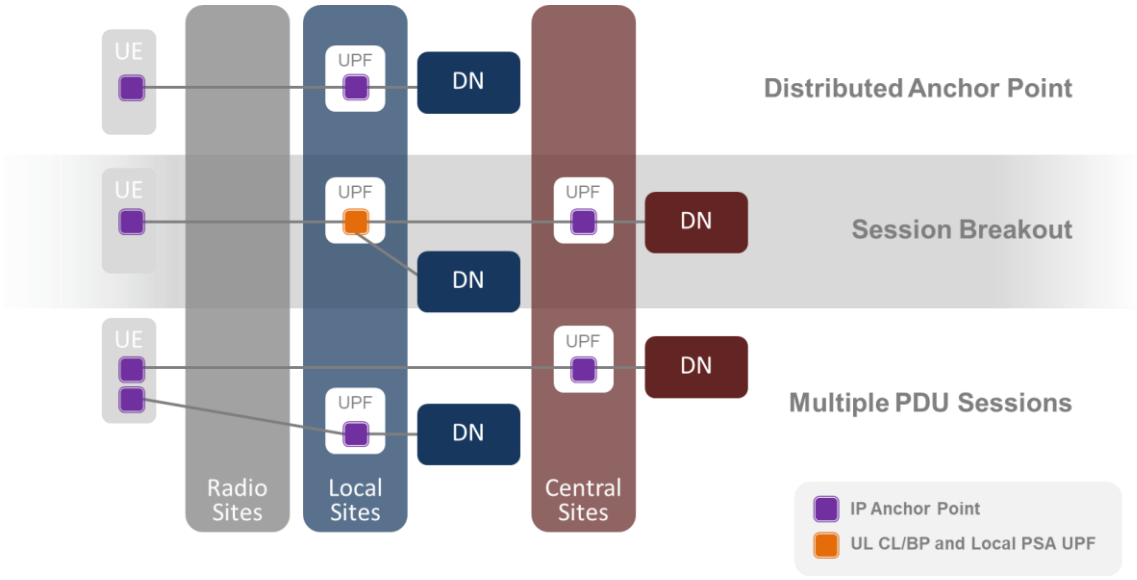

Figure 1. 5GC Connectivity Models for Edge Computing

This whitepaper explores various traffic engineering methods to address these challenges and optimize AI workload distribution to MEC resources in mobile network.

2 Deploying DN for Edge AI on network operator's site

Here are the key benefits of deploying DN for Edge AI on Local site than on deploying it on the Internet.

- Fine-grained traffic steering: It allows for precise control over traffic flows, enabling optimal routing of AI workloads to specific edge resources. Moreover, whole Edge AI traffic is closed within the carrier, meaning it is not exposed to the Internet, allowing for more reliable and flexible QoS control.
- Enhanced Privacy and security: Keeping sensitive AI data within a defined geographical area can help address data sovereignty and privacy concerns.
- Customized Services: Can be tailored to support specific AI applications or use cases within a given area, such as smart cities or industrial IoT deployments.
- Low Latency: DN's localized processing capability is ideal for AI applications requiring real-time responses, such as autonomous vehicles or augmented reality.
- Geographical Service Restriction: AI services can be confined to specific areas, which is useful for location-sensitive AI applications or to comply with data regulations.

On the other hand, here are the cons of deploying DN for Edge AI on Local site.

- Complexity: Managing multiple DNs and ensuring seamless integration with the broader network can increase operational complexity.
- Interoperability Concerns: Ensuring compatibility between DN-based AI services and non-DN network segments requires careful planning.
- Mobile network coverage dependency: It requires radio access network coverage, limiting its applicability in mobile network.

3 Traffic Engineering Methods for Edge AI Workloads

This section discusses three different traffic engineering methods for edge AI workloads, (1) Distributed Anchor Point, (2) Session Breakout, and (3) Multiple PDU Sessions. (3) has 3 more subcategories for traffic engineering. We review each method and provide pros and cons of each approach.

3.1 Distributed Anchor Point

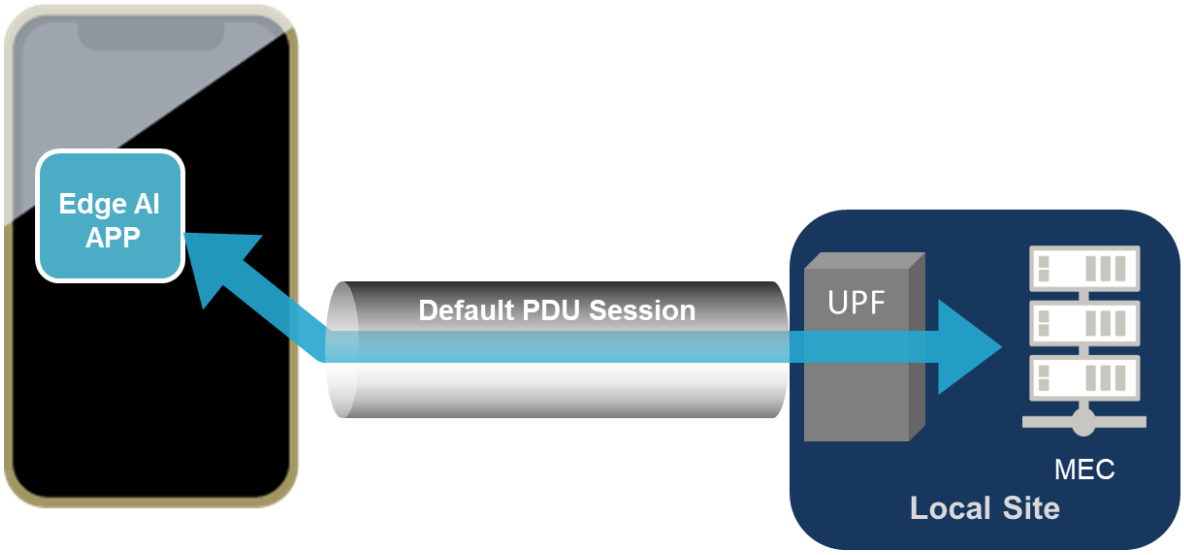

Figure 2. Distributed Anchor Point

This mechanism establishes a PDU session with the nearest Local Site (as illustrated in Figure 2). It is useful for single-function capability devices.

A UPF serving as an Anchor Point is placed in each region, and the UE connects to the nearest UPF by using the corresponding DNN. When the region changes, the network switches the connected UPF accordingly. This approach is suitable for nationwide service deployment rather than being restricted to a specific area. Nationwide services, meaning a service available throughout the entire country without being restricted to a specific area, are available by configuring a single DNN in the UE.

This configuration can be realized by leveraging standard 5G functionalities and does not require specific feature development in both UE and network.

It should be noted that for smart devices, up to 4G, PDU session for data communication always goes through the Central Site, and both the system and applications are structured accordingly. Therefore, considering interworking with 4G, this configuration is not suitable for smart devices as of now.

3.2 Session Breakout (ULCL)

ULCL is a feature introduced in 5G networks that provides a mechanism to classify and route uplink traffic (as described in Section 5.6.4.2 of TS23.501 [6]). A PDU session has a UPF towards the Central Site and another UPF (I-UPF) for local breakout (as illustrated in Figure 3). By combining this with PRA (Presence Reporting Area, as described in Section 5.6.11 of TS23.501 [6]), a service area which service is provided in specific area can be created. This capability is particularly valuable for directing AI-related data streams to appropriate edge computing resources. Since one PDU session is split into two, the security level of Local Site is the same as that of Central Site. Existing DNN configured in the UE can be used as it is, and no additional DNN is required.

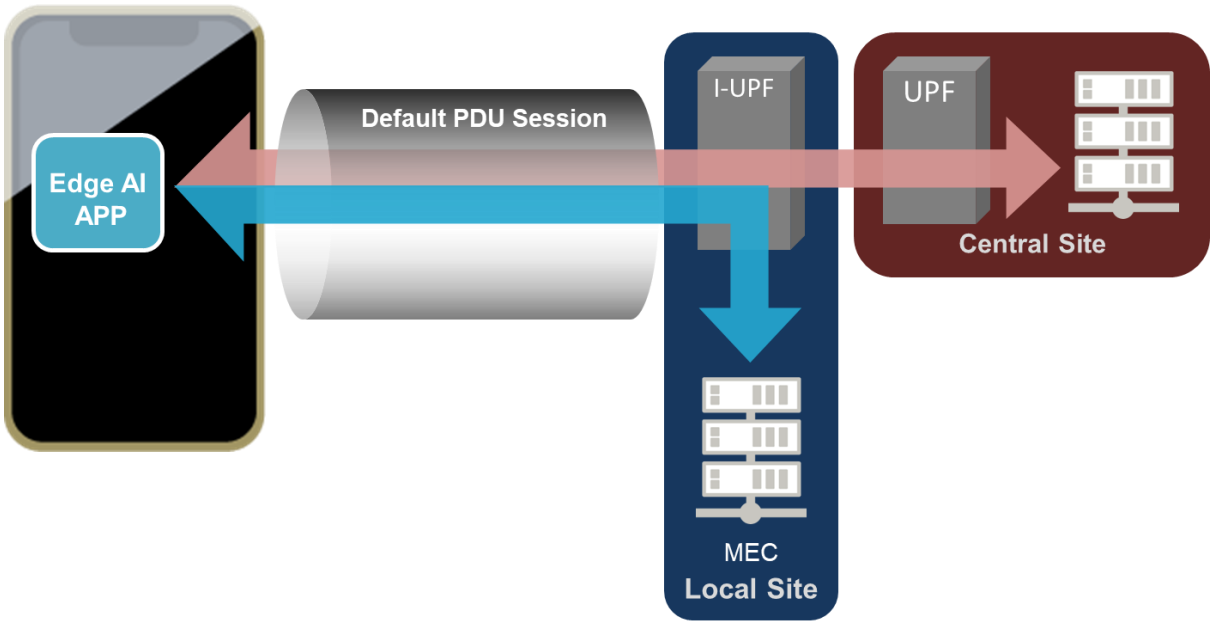

Figure 3. Session Breakout

Since the processing is completed within 5GC, the UE is unaware of the traffic diversion by the ULCL and does not participate in either the insertion or the removal of ULCL. No support is required on the UE side, making it possible to provide service to all UEs. It has already been confirmed that ULCL is applicable to all UEs.

Meanwhile, the UE's IP address is assigned by the UPF in Central Site, and the same IP address shall be used in Local Site. The network within Local Site is required to be designed carefully.

ULCL offers a tool for directing AI traffic in 5GC, enabling fine-grained control over how data is routed to edge computing resources. Its flexibility and adaptability make it well-suited to handle the diverse and evolving nature of AI workloads. However, the complexity of implementation within the mobile network and the potential for overhead must be carefully balanced against the benefits. As edge AI continues to grow in importance, ULCL is likely to play a crucial role in optimizing network performance and resource utilization for AI applications. Network operators and AI service providers will need to work closely to develop effective ULCL strategies that can fully leverage the capabilities of edge computing for AI workloads.

3.3 Multiple PDU sessions

To create dedicated PDU session for localized data networks is an advanced method for directing AI workloads to edge computing resources in Local Site (as illustrated in Figure 4). For the PDU session for Edge AI, a UPF that is geographically close to the MEC is selected. The DNN of this dedicated PDU session is different from that of the default PDU session. This approach leverages the flexibility and programmability of the 5GC network architecture to provide fine-grained control over traffic routing.

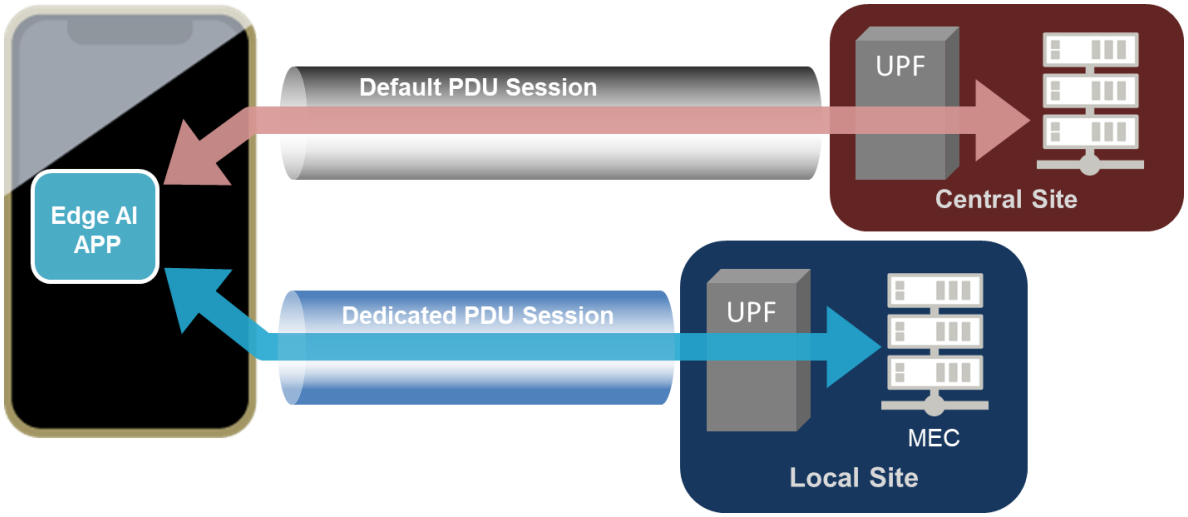

Figure 4. Multiple PDU sessions

For smart device such as smart phone or tablet, when two PDU sessions exist, there are two methods for associating PDU sessions on the UE side:

1. Application-Specific Selection: Each application selects and uses a specific PDU session. In this case, a single application can also utilize two PDU sessions simultaneously.
2. Instance-Based Segmentation: The device separates instances for Central Site and Local Site and associates each instance with a corresponding PDU session. This method assigns different PDU sessions for the Edge AI application and other applications using the Central Site. Since the instances are separated, it becomes easier to apply distinct security levels to each.

When two PDU sessions are used simultaneously, routing is required to associate applications with PDU

Sessions. This routing policy is implemented in UE Route Selection Policy (URSP) (described in Section 6.6.2 in TS 23.503 [8]) which can only be configured from the mobile network by over-the-air. In Android®[1] OS, slice categories such as "Latency" and "Enterprise" are available, and the network shall control how these are mapped to specific Data Network Names (DNNs) and/or Single Network Slice Selection Assistance Information (S-NSSAIs).

Since the dedicated PDU Session has been created, a different security policy can be applied to it and the network within the Local Site can be designed simply.

As specific examples of how to achieve local connectivity to a specific area using Multiple PDU Sessions, the following three methods are introduced.

3.3.1 Distributed Anchor Point for Dedicated PDU Session

This method creates a dedicated PDU session, apart from the default PDU session. In other words, it utilizes Distributed Anchor Point, introduced in Section 3.1, as the dedicated PDU session. It is useful for Smart Devices since there is a connectivity with Central Site using Default PDU Session, whereas Distributed Anchor Point is useful for single-function capability devices.

The functionalities required for this method by mobile network are the same as those needed for Distributed Anchor Point which can be implemented using standard 5G functionalities.

3.3.2 LADN

Local Area Data Network (LADN) is a feature introduced in 5G networks that enables the creation of localized data networks within specific geographical areas (as described in Section 5.6.5 of TS23.501 [6]). This approach is classified as a type of Multiple PDU sessions-based method. Instead of using Presence Reporting Area (PRA) (as described in Section 5.6.11 of TS23.501 [6]) to associate areas with PDU sessions, it utilizes the 3GPP standard’s specific LADN service area. It’s worth noting that an LADN service area is defined as a set of Tracking Areas. This capability is particularly relevant for directing AI traffic to MEC resources, as it allows for efficient handling of AI workloads within a defined zone.

The mobile network detects when the UE enters the designated area and instructs it to establish a PDU session. The UE does not need to monitor where the area is. When UE moves out of the area, the mobile network maintains the session for a certain period to avoid ping-pong effects at area boundaries. This eliminates unnecessary transactions. When the timer expires, the mobile network triggers the deletion of the session.

LADN is well suited to provide services in specific areas. LADN has already been implemented by a major

[1] Android is a trademark of Google LLC.

modem vendor and many Network Function vendors. The interoperability and effectiveness have already been confirmed by using off-the-shelf smartphone.

On the other hand, LADN requires a dedicated DNN for each area (note that dedicated S-NSSAI is not necessary). Therefore, it is not suitable for nationwide deployment.

3.3.3 Service Area Dedicated S-NSSAI

This method establishes a PDU session using a dedicated S-NSSAI and DNN that are available only within a specific area. When the UE enters the designated area, it utilizes these identifiers to set up the PDU session. Similar to LADN, a dedicated DNN and S-NSSAI is required for each area, making it unsuitable for nationwide deployment.

This approach may be realized by leveraging standard 5G functionalities and does not require specific feature development in the 5GC. However, UE shall monitor where the area is and trigger the establishment of the PDU Session. When the UE moves out of the area, the PDU session is deleted immediately. When a UE moves near area boundaries, it creates a ping-pong effect and generates unnecessary transactions. Therefore, it is not a better option than LADN even if service is provided in specific area.

3.4 Comparison

This section provides a comparison summary of the various traffic engineering methods for edge AI workloads discussed in the previous sections. Comparison Matrix of Breakout Mechanisms is shown in Table 1.

	Distributed Anchor Point	ULCL	Multiple PDU session		
			Distributed Anchor Point for Dedicated PDU session	LADN	Service Area Dedicated S-NSSAI
Local Site Connection for Area-specific service (*1)		✓		✓	✓
Local Site Connection for Nationwide service (*2)	✓		✓		
Connection unit to Local Site	PDU session	IP flow	PDU session	PDU session	PDU session
Interworking with 4G		✓	✓	✓	✓
Feasibility of security enhancement	✓		✓	✓	✓
Network design complexity	Low	High	Low	Low	Low
Frequency of signaling involving UE at area boundaries	High	None	High	Low	High
Implementation impact on UE	None	None	None	Low (Supported by major vendors)	Moderate
Implementation impact on 5GC	None	High	None	Low (Supported by some vendors)	Low

Table 1. Comparison Matrix of Breakout Mechanisms

(*1) “Local Site Connection for Area-specific service” means connecting to the Local Site in specific locations such as particular offices or factories. If the UE leaves the designated area, it will lose connection to the Local Site.

(*2) “Local Site Connection for Nationwide service” means establishing a Local Site connection for each region where UE is located. For example, if a UE is in Region A, it will connect to the Local Site of Region A; if a UE is in Region B, it will connect to the Local Site of Region B.

Key Findings:

1. ULCL is suitable for services provided to mass users since it does not have UE dependency.
2. LADN and Service Area Dedicated S-NSSAI can separate UE instances for Local Site connection, thus enhancing security for MEC connections. Therefore, they are suitable for services provided by corporations to their self-managed UEs.
3. ULCL, LADN, and Service Area Dedicated S-NSSAI are designed to provide services within a specific area.

Distributed Anchor Point is a more practical approach for services that are not restricted to a specific area.

4. It is important to note that these methods are not exclusive. These methods can co-exist within a UE by using different applications for each method.

4 Edge Service Enablement and Deployment

4.1 Service Location Management

Edge services are distributed across multiple locations within the operator's service network. The same service may be implemented in various locations. When accessing services implemented at the edge, user applications need to identify the most suitable service location. For instance, if communication latency is the most critical factor for an application, the closest edge location should be selected. Conversely, if power consumption is a priority, the application may choose a more distant but cost-effective data center.

Service locations are ultimately bound to network addresses. To provide optimal locations to applications, we need a mechanism to map user application requirements to service locations. This mapping can be implemented using the DNS system. When we receive a service location query from an application, we consider factors such as the application type and the query's source location to map the best edge location to the query. Additional parameters may be considered when mapping services to locations.

Once the application receives the service location, which is an IP address, it initiates communication with the given destination address.

4.2 Interworking with UEs under non-AI-RAN Coverage

The concept of AI-RAN is to share common infrastructure for both software-based RAN and AI services. This integrated approach means that the closer these functions are deployed together, the more efficiently we can route RAN communication to AI services, thereby reducing latency and enhancing performance. The ultimate vision for AI-RAN is that every RAN access point will be equipped with AI-RAN capabilities and MEC functions, creating a seamless and efficient network landscape.

However, the deployment of AI-RAN will not happen overnight but rather will occur incrementally. Initially, only a limited number of areas will be included in the AI-RAN infrastructure, while the rest will continue using traditional RAN service models. In these early stages, the full benefits of AI-RAN, such as infrastructure sharing and optimized resource usage, will not be fully realized. Nonetheless, these areas can still leverage MEC functionalities in much the same way as traditional MEC sites. The technologies discussed in this white paper are applicable to traditional RAN services as well. The key difference is that under traditional models, RAN services and AI services are provided by separate entities. Even in this scenario, AI services can still be delivered to existing traditional RAN users through the partially deployed AI-RAN infrastructure.

One issue to note is that since RAN processing and AI processing are performed on different equipment, some of the more advanced MEC services that require integration of both RAN software information and AI service data may not be available. However, many AI/MEC services can still be implemented.

Telecommunications operators can strategically plan and incrementally deploy AI-RAN infrastructure within their service networks, tailoring their approach based on their investment strategies and long-term goals. This phased deployment allows operators to manage costs while gradually enhancing their service capabilities, laying the groundwork for a future where fully integrated AI-RAN networks provide cutting-edge services and superior performance.

5 AITRAS

AITRAS is a SoftBank version of AI-RAN implementation, build on top of GPU-based accelerator hardware. We are planning to evaluate what is the feasible AI-RAN design to our customers, including consumer customers and enterprise customers. Since no LBO technology fits for all the situation, we may need to prepare options for different kinds of LBO service architecture depending on the service requirements. In this section, we introduce two use cases that can be implemented based on the AITRAS concept: an LLM service using RAG and an LLM service for confidential information. Since AITRAS is a versatile infrastructure providing powerful software capabilities, including RAN and AI, we can envision many other applications as well.

5.1 AITRAS Use Cases

As outlined in this document, there are multiple methods for implementing local breakout. In this instance, we have realized two use cases utilizing the LADN and URSP functionalities:

1. LLM switching application utilizing URSP
2. Multimodal RAG application utilizing LADN

Use Case 1: LLM switching application utilizing URSP

In this use case, confidential and general information are processed through an agent on the smartphone, where confidential information is directed to a closed network and general-purpose information is directed to an open cloud environment. This use case necessitates establishing two PDU sessions from a single application, which we have validated using URSP (as illustrated in Figure 5). This use case uses Distributed Anchor Point for Dedicated PDU session introduced in Section 3.3.1.

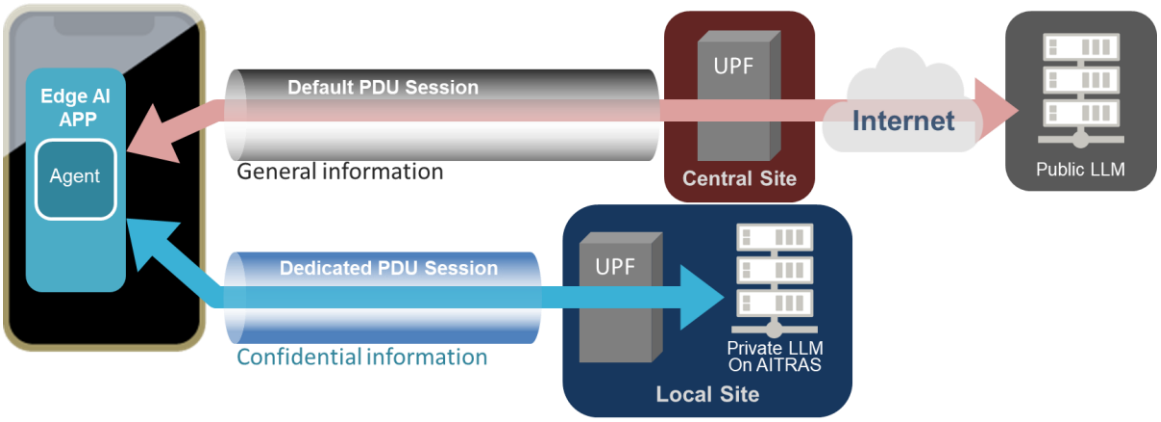

Figure 5. LLM switching application utilizing URSP

Use Case 2: Multimodal RAG application utilizing LADN

This use case leverages LADN to achieve local breakout, thereby accessing the AITRAS resources (as illustrated in Figure 6). Specifically, in the factory use case, high security is required; and there is an anticipated need to allow access only to users within the factory thus, utilizing LADN technology for access is deemed appropriate.

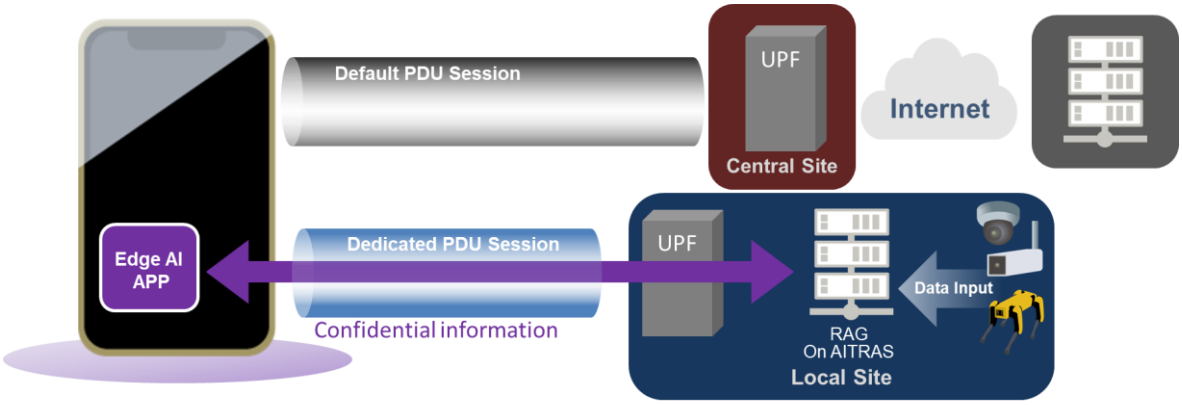

Figure 6. Multimodal RAG application utilizing LADN

For further details, please refer to the press release at the following URL:
[https://www.softbank.jp/en/corp/news/press/sbkk/2025/20250303_04/]

6 Conclusion

As we conclude this comprehensive exploration of directing AI traffic to Multi-access Edge Computing (MEC)

resources, it is evident that the landscape of edge AI is rapidly evolving, presenting both significant opportunities and complex challenges.

Our key findings reveal that the rapid growth of AI services is driving unprecedented demand for edge computing resources, necessitating sophisticated traffic engineering methods. Various approaches, including Distributed Anchor Point, Session Breakout, Multiple PDU Sessions, offer unique advantages and trade-offs considering operation management, service area definition, scalability, and implementation complexity. Integrating edge AI with existing non-AI-RAN radio UEs are possible, although some of the advanced integrated functions of RAN processing and AI processing may not be provided to traditional UEs. We observe that emerging technologies like 6G, advanced AI models at the edge, and autonomous network management are poised to further revolutionize edge AI traffic engineering.

Our insights highlight that due to the diverse nature of AI applications and network environments, no single traffic engineering solution fits all scenarios. A combination of methods may be necessary for optimal effectiveness. This requires network operators and service providers to carefully balance the need for advanced AI capabilities with the requirements of existing infrastructure and non-AI traffic. Additionally, ensuring robust security and privacy measures at the edge is crucial, given the sensitive data increasingly involved in AI workloads. The rapid pace of technological advancement in AI and networking necessitates flexible and adaptable approaches to traffic engineering.

We can expect to see a gradual blurring of the boundaries between edge and cloud computing. This evolution will likely result in a more seamless integration of edge and cloud resources, enabling fluid workload distribution across the entire compute continuum. Such integration will offer unprecedented flexibility in managing AI tasks, allowing for optimal resource allocation based on the specific requirements of each application.

Furthermore, the maturation of edge AI technologies will likely catalyze industry-wide efforts towards standardization. We can anticipate increased collaboration among stakeholders to develop comprehensive standards that ensure interoperability, efficient resource utilization, and consistent performance across diverse edge AI environments. These standardization efforts will be crucial in fostering a robust and scalable ecosystem for edge AI applications.

In essence, the future outlook for edge AI traffic engineering is one of increasing sophistication, integration, and efficiency. As these technologies continue to evolve, they will undoubtedly open up new possibilities for AI applications across various industries, potentially revolutionizing fields such as autonomous vehicles, smart cities, industrial IoT, and beyond.

The successful implementation of edge AI traffic engineering strategies will be crucial in unlocking the full potential of AI technologies across various industries. As we stand at the head of a new era in computing, the ability to efficiently direct and manage AI workloads at the edge will be a key differentiator for organizations seeking to leverage the power of AI.

The path forward will require continuous innovation, collaboration across sectors, and a commitment to addressing the technical, operational, and ethical challenges that lie ahead. By embracing these challenges and working together towards common standards and best practices, we can create a robust, efficient, and sustainable edge AI ecosystem that drives innovation and delivers value to users worldwide.

SoftBank will continue to work on AI-RAN and Edge AI processing through our research and development activities. AITRAS is one of our activities that emerges a part of our vision. We are committed to advancing traditional telecommunication business to AI service infrastructure provider through advanced networking operation technologies.

References

[1] AI-RAN Alliance, "https://ai-ran.org/," [Online].

[2] AI-RAN Alliance, "AI-RAN Alliance Vision and Mission White Paper," [Online]. Available: https://ai-ran.org/wp-content/uploads/2024/12/AI-RAN_Alliance_Whitepaper.pdf.

[3] L. Kundu, X. Lin, R. Gadiyar, J.-F. Lacasse and S. Chowdhury, "AI-RAN: Transforming RAN with AI-driven Computing Infrastructure," arXiv.

[4] SoftBank Corp., "AI-RAN: Telecom Infrastructure for the Age of AI," [Online]. Available: https://www.softbank.jp/en/corp/technology/research/news/061/.

[5] SoftBank Corp., "Telco AI: Landscape, Challenges, and Path Forward," [Online]. Available: https://www.softbank.jp/en/corp/technology/research/news/064/.

[6] 3GPP, "TS23.501: System architecture for the 5G System (5GS)," 3GPP.

[7] 3GPP, "TS23.548: 5G System Enhancements for Edge Computing," 3GPP.

[8] 3GPP, "TS 23.503: Policy and charging control framework for the 5G System (5GS)".

Acknowledgment

This white paper was authored by SoftBank's Research Institute of Advanced Technology team, with contributions from Ryuji Wakikawa, Keiichi Shima, Keisuke Asakura, Noriaki Yamada, and Tsuneomi Haruna.

The Research Institute of Advanced Technology at SoftBank, established in April 2022, is committed to advancing cutting-edge technologies that drive societal progress. As an *activator* for innovation, it focuses on research in AI-RAN, 6G, HAPS (High Altitude Platform Station), autonomous driving, quantum technologies, and other forward-looking fields.

Our mission is to harness the power of technology to create solutions that address real-world challenges and shape a better future. Guided by SoftBank's corporate philosophy, "Information Revolution - Happiness for everyone," we strive to lead the next wave of advancements that improve lives and inspire progress.

For more details about our latest technological developments and initiatives, please visit our website: https://www.softbank.jp/en/corp/technology/research/