

AI

RAN

White Paper

AI-RAN Orchestration: A Practical Step Toward Shared Infrastructure

— Architecture, trade-offs, and considerations —

July 2026
SoftBank Corp.

 SoftBank


ERICSSON

AI-RAN オーケストレーション: 共有インフラへの実践的な一歩

アーキテクチャ、トレードオフ、検討事項

ソフトバンク | Ericsson

はじめに

没入型体験からリアルタイム意思決定システムに至るまで、AI駆動型アプリケーションは、レイテンシ、アップリンク容量、ならびに無線・伝送・コンピューティング全体にわたるオーケストレーションに対して、新たな要件を突きつけています。同時に、将来のネットワーク運用では、インテリジェントなネットワーク最適化、動的スライシング、エネルギー効率化、ライフサイクル管理など、より高度な自動化が必要になります。こうした自動化は、ネットワーク制御とコンピューティングリソースをより緊密に連携させることで効果を発揮するAI技術への依存を強めています。この2つの動きにより、RANの役割は変化しつつあります。RANは、インテリジェントなネットワークの一部であるだけでなく、AIを実行するための重要な基盤としても位置づけられるようになっていきます。

本稿では、RANとAIが共有アーキテクチャフレームワーク内でどのように共存できるかを検討します。また、新たに登場するAI駆動型ユースケースを支えるために、ネットワーク側にどのような機能が必要になるかも扱います。議論の中心となるのは、RAN機能とAIワークロードを共存させるためのアーキテクチャ上の考慮事項です。これには、自動化、オーケストレーション、ライフサイクル管理への影響も含まれます。rAppは、RANの自動化機能を柔軟にデプロイし、独立して管理するための仕組みです。そのため本稿では、検討対象となるAI-RANワークロードを、RANベースバンド機能やその他のリアルタイムRANワークロードではなく、rApp[1]として実行されるワークロードと想定します。



この前提のもと、本稿では、ソフトバンクとEricssonが共同で開発した具体的なオーケストレーションのユースケースを取り上げます。このユースケースは、ソフトバンクのAITRAS OrchestratorとEricsson Intelligent Automation Platform(EIAP)が連携し、rAppsを含むAIワークロードの計算需要に応じてリソース割り当てを動的に調整することで、AI-RAN Orchestrationの可能性を示すものです。

この連携により、AIワークロードとRANインフラが協調して動作するAI-and-RANのユースケースが実現されます。また、こうした協調は、AI-on-RANワークロードのさらなる活用に加え、ネットワーク自体における追加的な自動化および最適化機能、すなわちAI-for-RANの実現にもつながります。

AI-and-RANオーケストレーション

AI-and-RANの概念は、同一のITインフラ上でAIワークロードとRANワークロードの双方を実行し、インフラのリソース利用率と柔軟性を高める機会を生み出すものです。

コンシューマ向けおよびエンタープライズ向けのAIワークロードは、通常、性質の異なる実行フェーズに分かれます。トレーニング、再トレーニング、ライフサイクル管理は一般的に集中型の環境で実行され、大量のデータを高速に処理し、GPUなどの計算資源を効率よく使うことが重視されます。一方、推論ワークロードはエッジに近い場所で動作し、バッチ処理から、低遅延が求められるインタラクティブサービスまで、用途に応じて幅広い形態を取ります。

同様に、RANワークロードも、タイミングや信頼性に関する要件が様々ではありません。リアルタイム領域で動作する時間制約の厳しい処理から、ベースバンド処理経路の外側で動作する比較的柔軟な機能まで、幅広い種類の機能が含まれます。一部のRAN機能は厳格なレイテンシ制約を必要としますが、その他の機能はある程度の変動を許容できるため、非リアルタイム環境での実行にも適しています。

RANワークロードの中でも、rAppはRAN自動化に用いられるワークロードの一種ですが、その実行特性、特に機械学習モデルに関連する特性は、時間制約の厳しいRAN処理よりもAIワークロードに近いものです。これらのモデルは、多くの場合、トレーニングまたは再トレーニングのフェーズで一時的に多くのコンピューティングリソースを必要とし、定常運用中には必要なリソースが低下します。



このため、rAppは、同一インフラ上でAIワークロードとRAN機能がどのように共存できるかを検討するための具体例となります。この点については次章で詳述します。

リソース管理シナリオとアーキテクチャ上の影響

異なる実行特性を持つAIワークロードとRANワークロードを同じインフラ上で共存させる場合、リソース管理の方法は、どのレベルでインフラを共有するかによって変わります。共有のレベルには複数の段階があり、ワークロード間の結合度が高まるほど、リソース利用の柔軟性は高まります。一方で、ワークロード同士をどの程度分離できるか、運用がどれだけ複雑になるかといった点では、トレードオフが生じます。

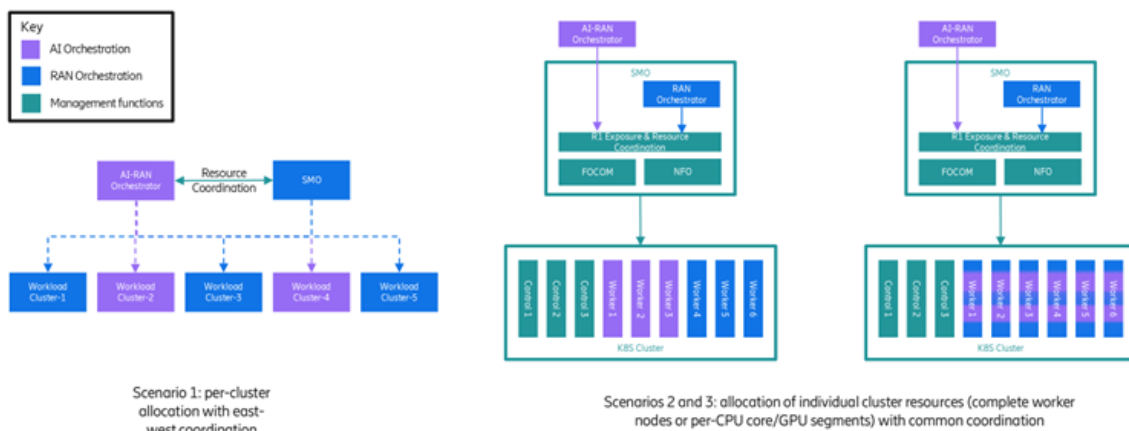


図1: AIとRANの共有シナリオ

代表的なシナリオとして、次の3つが考えられます。

- シナリオ1: 複数クラスター

RAN用、AI用といった形で、ワークロードドメインごとに個別のクラスターを割り当てます。この構成では、各ワークロードは別々のクラスター上で実行されます。

- シナリオ2: 共有クラスター

単一のクラスターを、RANワークロードとAIワークロードで共有します。ただし、ワークロード同士は、より高い抽象レベルで分離されます。



- シナリオ3:共有サーバー

同一サーバー上で、RANワークロードとAIワークロードがリソースを共有します。

3つのシナリオの中で、最も細かい粒度でリソースを共有する構成です。

これらのシナリオは、ワークロードを明確に分離する構成から、より緊密に統合する構成へと段階的に進むものです。共有の粒度が細くなるほど、リソース利用率を高められる可能性は大きくなります。一方で、ワークロード間の調整、性能保証、障害分離に対する要求も高まります。

以下の表は、これらのトレードオフを示しています。

観点	シナリオ1:複数クラスター	シナリオ2:共有クラスター	シナリオ3:共有サーバー
リソース共有のレベル	クラスター間の共有なし	クラスターレベル	サーバーレベル
ワークロード分離	非常に高い	高い	限定的
運用複雑性	低い	中程度	高い
性能予測可能性	高い	中程度	より低い
ワークロード間干渉のリスク	最小限	限定的	より高い
障害分離	明確	おおむね局所化	より複雑
リソース効率	低い	中程度	高い
現行デプロイ方式との整合性	強い	部分的	限定的
ビジネスリスク	低い	中程度	高い



現行のデプロイ方式との整合性

特定のワークロードドメインごとに専用のクラスターをデプロイする方法(シナリオ1)は、現在の商用ネットワークにおけるRANおよび周辺コンピューティングリソースのデプロイ方式とよく整合しています。

このアプローチでは、ワークロード間の分離境界が明確になり、性能も予測しやすくなります。また、障害が発生した場合にも影響範囲を限定しやすくなります。これらはいずれも、ネットワークの信頼性を維持するうえで重要です。

この方法は、インフラリソースの共有を最大化するものではありません。しかし、異なる種類のワークロードをより緊密に結合することを正当化できる明確なビジネスケースがない場合には、不要な複雑性を避けられるという利点があります。

運用面では、このように分離されていることで、各ワークロードドメインは、それぞれの要件に応じたオーケストレーション機構やライフサイクル管理プロセスを用いて、独立して発展させることができます。同時に、必要に応じてインフラレベルで協調することも可能です。

共有クラスターおよび共有サーバーに関する検討事項

単一のクラスターまたは単一のサーバーをAIとRANのワークロードで共有する方式(シナリオ2および3)は、柔軟性やリソース利用率の向上が期待できる一方で、ワークロードドメイン間の結合がより緊密になります。こうしたシナリオでは、特にワークロードごとに時間的要件やスケーリング特性が異なる場合、リソース競合、障害波及、性能面での相互作用を把握・管理することがより困難になります。

強力かつ明確なビジネス上の目的がない限り、このような複雑さは、より細かな粒度でのリソース共有によるメリットを上回る可能性があります。特に、RAN性能の低下や運用負荷の増大につながる場合はなおさらです。そのため、共有クラスターおよび共有サーバー方式は、協調メカニズムの成熟や、ビジネス上の価値の明確化が進むことで、将来的に有力な選択肢になると考えられます。



各シナリオにおけるオーケストレーション上の影響

シナリオ1では、各ワークロードドメインに専用のクラスターと、それに対応するオーケストレーション機能が割り当てられます。そのため、AIソフトウェアとRANソフトウェアは、それぞれ独立してオーケストレーションできます。

インフラ共有が必要な場合には、各ドメインの需要変動に応じて、オーケストレーターがクラスター間でリソースを要求したり解放したりすることで対応できます。これにより、各ドメインの所有境界を明確に保ちながら、インフラレベルでは制御された形でリソースを共有できます。このシナリオについては、後続の章でより詳細に検討します。

シナリオ2およびシナリオ3では、ワークロードが共有クラスター内、または共有サーバー上に配置されます。そのため、シナリオ1よりもワークロード間の緊密な調整が必要になります。

これらの場合、複数のオーケストレーション主体がリソースを扱うため、競合や意図しない相互作用を避ける必要があります。特に重要なのは、複数のオーケストレーターが同じリソースに対して、それぞれ独立に制御権を持つとみなしてしまう「マルチマスター」状態を回避することです。

考えられるアーキテクチャ上の方向性の1つは、AIおよびRANワークロードドメイン固有のオーケストレーションロジックと、Kubernetesクラスター内のリソース割り当てを調整する汎用機能を分離することです。これにより、各ワークロードに固有の要件は独立して扱いながら、クラスター・レベルのリソース管理については共通の調整ポイントに集約できます。

このようなアプローチは、O-RAN AllianceのService Management and Orchestration (SMO) フレームワークのようなアーキテクチャと概念的に整合しています。同フレームワークでは、ワークロードをオーケストレーションする機能と、標準化インタフェースを通じて基盤インフラへの変更を調整する仕組みが区別されています。



AI-and-RAN: AI-RAN rAppと非RANワークロードによるエッジリソース共有の実証

現行のデプロイ方式との整合性を踏まえると、AI-and-RANの初期的な実現形態としては、ワークロードドメイン間で緩やかに連携しながらコンピューティングリソースを共有する方法が考えられます。

異なる種類のワークロードを細かい粒度で同じ場所に配置するのではなく、明確な所有権と運用上の分離を維持しながら、ドメインをまたいでAIに最適化されたコンピューティングリソースを一時的かつ制御された形で利用できるようにすることが、より現実的なアプローチです。

代表的な例の1つは、通常はAIドメインで利用されるGPU搭載のクラウドベース・コンピューティングリソースを、オフピークなどリソース利用率が低い時間帯に、RAN関連のAIワークロード向けに割り当てることです。これにより、AIモデルのトレーニングまたは再トレーニングを、RANドメイン内にリソースを恒久的に確保することなく実行できます。また、AIとRANの実行環境を過度に緊密に結合することも避けられます。

Ericssonとソフトバンクは、このようなユースケースを実証するために協力し、AI-for-RAN環境とAI-on-RAN環境を組み合わせたAI-and-RANシナリオを示しました。このシナリオは、上記のような緩やかなリソース共有モデルに整合するものです。

ソフトバンクのラボ環境では、複数のシングルノードクラスターがAI-RANオーケストレーターの下に置かれています。一方、RANワークロードはこの実証のためにシミュレートされ、EricssonラボにおいてSMO、具体的にはEricsson Intelligent Automation Platform (EIAP) の制御下で実行されます。

この構成では、rAppがRANワークロードとして機能し、最適化関連機能を実行します。また、既存のRAN運用原則を損なうことなく、AI関連のRANワークロードが共有コンピューティングリソースをどのように活用できるかを示す具体例となります。



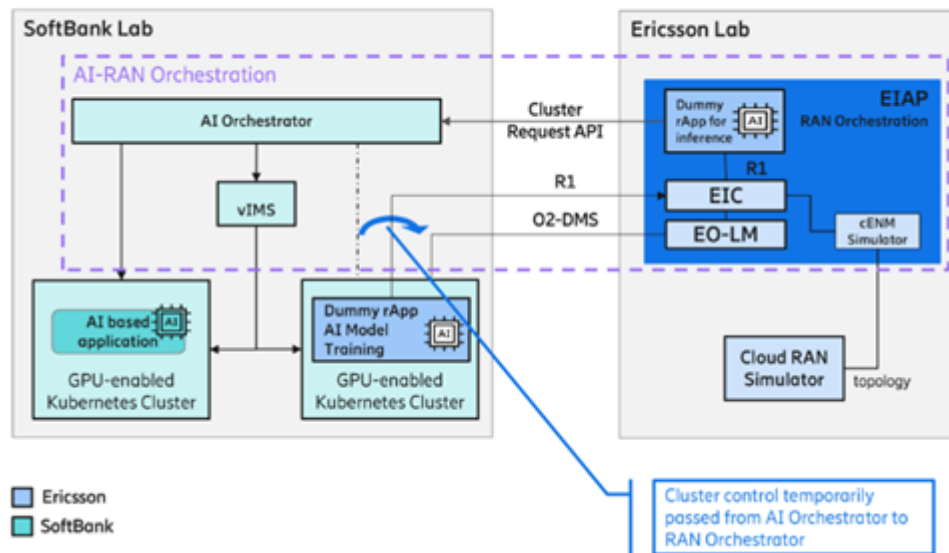


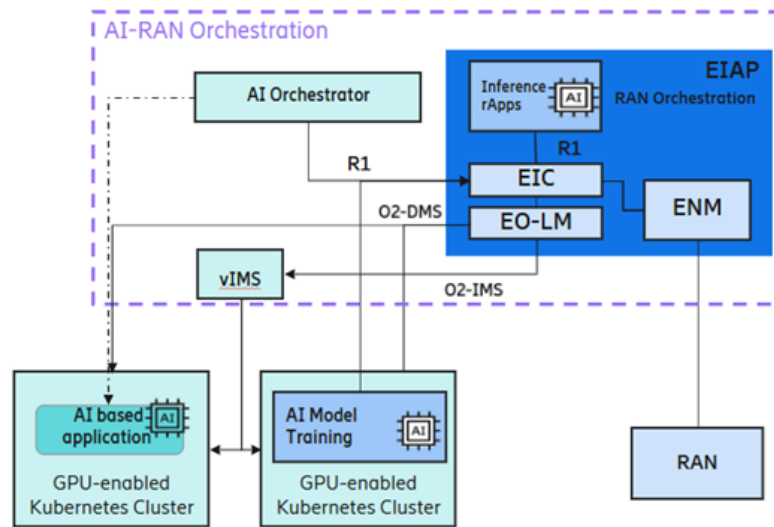
図2:AI-RANラボ実証構成

通常運用時、rAppはEricssonラボ内のSMO環境にある、汎用CPUベースのKubernetesクラスター上でAIモデルを実行します。一方、ソフトバンクのラボ環境にあるAI-RANオーケストレーターは、ソフトバンクのGPUベースのインフラ上で、エンタープライズユーザー向けのAIワークロードを独立して管理しています。また、AI-RANオーケストレーターは、EIAPから追加リソースを要求できるようにするため、R1拡張APIをEIAPに登録しています。

ある時点で、EIAP環境内で稼働しているrAppは、自身のモデル品質が低下しており、新しいデータを使った再トレーニングが必要であることを検知します。そこでrAppは、R1経由で登録されたAI-RANオーケストレーターAPIを通じて、新たなネットワークデータを用いたトレーニングセッションを実行するために、GPU対応Kubernetesクラスターへのアクセスを要求します。

AI-RANオーケストレーターは、利用可能なクラスターのインベントリーを確認し、現在どのAIアプリケーションにも使用されていないクラスターがあるかを判断します。利用可能なクラスターがある場合、AI-RANオーケストレーターはそのクラスターを予約済みとしてマークし、再トレーニングを必要としているrAppに接続先などの詳細情報を渡します。AI-RANオーケストレーターは、時間ベースのポリシー、将来負荷の予測、その他の運用上の条件に基づいて、クラスターを利用可能にするかどうかを判断できます。





rApplは、Federated O-Cloud Orchestration and Management(FOCOM) APIをR1経由で使用し、この新しいクラスターをワークロードクラスターとしてSMOに登録します。その後、rApplはNetwork Function Orchestration(NFO)APIをR1経由で使用して、モデル・トレーニングワークロードをデプロイできます。また、トレーニングに必要なデータセットへの安全なアクセスは、R1経由のData Management and Exposure(DME)APIを通じてワークロードに提供されます。

モデル・トレーニングワークロードは、クラスターのGPUを活用し、多数のトレーニングエポックを高速に実行します。これにより、モデル精度を段階的に向上させることができます。

その後、モデルが十分にトレーニングされるか、AI-RANオーケストレーターがクラスターの返還を要求した時点で、トレーニングワークロードは終了されます。終了後、更新済みモデルはR1 DME APIを介してrAppから利用できるようになります。

このユースケースは、以下を実証するための実践的な第一歩です。

- RAN関連ワークロードとその他のエッジAIアプリケーションの間でコンピューティングリソースを動的に割り当てることで、ハードウェアをより効率的に活用し、新たな収益機会を創出できること。
- rAppのAIモデルと非RAN AIワークロードが、制御された形でエッジインフラを共有できること。

- EricssonのSMOとソフトバンクのAI-RAN Platform and Infrastructure Orchestratorを統合しO-RANアーキテクチャ内のR1インタフェースを活用することで、疎結合の形で他のワークロードオーケストレーターと共存・協調できること。
- 共有インフラ上での協調的なオーケストレーションを支援するために、R1経由でNFOおよびFOCOMインタフェースを公開できること。これにより、O-RANアーキテクチャの拡張性を示せること。

なお、本ホワイトペーパーで議論・検討しているアーキテクチャは、AI-RAN Allianceで定義されているAI-RANオーケストレーションのリファレンスアーキテクチャ [2] に整合するものです。また、ソフトバンクのAITRAS Orchestratorは、[3] で定義されているAI-RAN Platform and Infrastructure Orchestratorに対応するものです。

今後の考慮事項: ビジネスおよび技術の観点

AIワークロードとRANワークロードのインフラレベルでの連携が進むにつれて、ビジネス面と技術面の双方で、検討すべき課題が残されています。

技術の観点では、実行特性の異なるワークロードを、予測可能な性能と障害分離を維持しながら、どのように共存させるかをさらに検討する必要があります。そのためには、リソース抽象化、ワークロード特性の把握、インタフェースの進化が、不要な結合や運用リスクを招くことなく、より柔軟なデプロイをどのように支援できるかを明らかにする必要があります。

これらの検討事項は、今後のAI-and-RANアーキテクチャの進化を方向づけるものです。その進化は、実際の運用経験と、明確なビジネス上の合理性の双方に基づいて進める必要があります。

本取り組みから浮かび上がる主な検討事項は、以下のとおりです。

ビジネス観点の検討事項

- リソース利用率の向上だけでなく、タイム・トゥ・マーケットの短縮やサービス差別化など、インフラ共有によって得られる具体的な価値を明確にすること。



- アーキテクチャの複雑化が、明確で定量的なビジネス上・経済上の合理性に見合っていることを確認すること。
- ワークロードドメイン間で、責任範囲と説明責任を明確に分離して維持すること。

技術観点の検討事項

- 性能を予測しやすい状態に保ち、障害発生時の影響範囲を分離できることを、重要な設計要件として維持すること。
- ワークロードの特性を把握し、配置先やリソース割り当てを適切に判断できるようにすること。
- 実証結果とビジネスニーズに基づき、ワークロード間の結合度を段階的に高めながら、アーキテクチャを進化させられるようにすること。

これらのビジネス面および技術面の検討事項は、アーキテクチャ上の選択を、実際の運用上の要件と、実証可能な価値の双方に整合させることの重要性を示しています。

AI-RANアーキテクチャが進化していく中で、このバランスを維持することは、持続可能でスケーラブル、かつキャリアグレード環境に適した共存モデルを実現するうえで重要です。

参考文献

1. rAppとは、Ericsson Intelligent Controller(EIC)上にデプロイされるRANアプリケーションです。EICはEricssonのNon-Real-Time RICであり、rAppはデータ、モデル、制御インタフェースを活用して、時間制約の厳しい処理の外側でRANの動作を最適化・自動化します。
2. AI-RAN Alliance、「AI-RAN Architecture v1.2」、2026年2月。
入手先:<https://ai-ran.org/documents/AI-RAN-Architecture-v1.2.pdf>
3. AI-RAN Alliance WG2(AI-and-RAN)、「AI-RAN Platform and Infrastructure Orchestrator」、2026年2月。



入手先:<https://ai-ran.org/documents/AI-RAN-Platform-Infrastructure-Orchestrator.pdf>

ソフトバンクについて

ソフトバンク株式会社は、「情報革命で人々を幸せに」というソフトバンクグループの経営理念の下、通信サービスと多様なソリューションを提供しています。

当社グループは、成長戦略「Activate AI for Society」を掲げ、全ての事業でAIの可能性を起動させ、AIの社会実装を推進することで企業価値の最大化を目指しています。また、AI計算基盤やAIデータセンターなどを通じて、AI時代の社会インフラの構築を推進しています。

Ericssonについて

Ericssonの高性能ネットワークは、毎日、世界中の数十億人に接続性を提供しています。Ericssonは約150年にわたり、通信技術の創出を先導してきました。サービスプロバイダーおよび企業向けに、モバイル通信とコネクティビティのソリューションを提供しています。お客様やパートナーとともに、明日のデジタル世界を実現します。

本書の内容は、方法論、設計、製造における継続的な進展により、予告なく改訂される場合があります。本書の使用に起因するいかなる誤りまたは損害についても、ソフトバンクは一切の責任を負いません。

© 2026 ソフトバンク株式会社



AI-RAN



 SoftBank


ERICSSON