THE INSIGHT RESEARCH CORPORATION

cohere technologies

Whitepaper – Cohere Technologies and the Deft Dance of the USM Delivery

Insight Research Corporation

August 2025

The contents of this study represent our analysis of the information generally available to the public or released by responsible individuals in the companies mentioned. It does not contain information provided in confidence by our clients. Since much of the information in the study is based on a variety of sources that we deem to be reliable, including subjective estimates and analyst opinion, Techgradient does not guarantee the accuracy of the contents and assumes no liability for inaccurate source materials. Techgradient disclaims liability for incidental or consequential damages and assumes no responsibility or liability for any loss or damage suffered by the buyer as a result of the use or misuse of any of the information or content of this study. Techgradient assumes or undertakes no liability for any loss or damage suffered as a result of the use, misuse or reliance on the information and content of this study.

Table of Contents

Tables and Figures

1. Executive Summary

The whitepaper examines three delivery modes for the Universal Spectrum Multiplier (USM) offered by Cohere Technologies (Cohere):

- xApp-based deployment via the Near-RT RIC - offers architectural elegance, vendor neutrality, and cloud-native scalability. It faces resistance due to business and political realities. Major RAN vendors, wary of losing control, resist third-party applications that attack the very core of their competencies. Despite successful PoCs with operators like Vodafone, widespread adoption remains constrained by OEM reluctance.
- Direct-to-base station integration - is a pragmatic middle ground. Using gRPC wrappers and Cohere's emulation framework, it enables quick deployment with minimal architectural disruption, especially in the vast installed base of FDD base stations. Cohere's work with Bell Canada illustrates is a case in point.
- Joint scheduling with embedded or dApp-assisted logic - represents the deepest integration, embedding USM logic within the base station MAC layer or using distributed applications (dApps). This approach requires full vendor cooperation, which is a long shot.

On the whole, the USM delivers a compelling response to the long-standing challenge of multi-user MIMO (MU-MIMO) adoption in RAN deployments. It leverages the Delay-Doppler domain— distinct from the conventional time-frequency model. USM improves spectral efficiency by dynamically optimizing user pairing and precoding. Trials have shown gains of up to 50% in spectral performance.

2. USM and the Delivery Dilemma

This section provides an overview of the value addition that USM brings to the table and lays ground for the comparisons of the diverse delivery mechanisms.

Cohere's USM exemplifies agility in addressing spectral efficiency through innovative deployment strategies in the RAN ecosystem. This whitepaper explores the technological and strategic evolution of USM and its three delivery modes:

- xApp,
- direct-to-base station, and
- joint scheduling

With proven spectral efficiency gains of up to 50%, USM exemplifies a promising yet complex, solution navigating a consolidated vendor landscape and telco conservatism. The paper provides a comparative analysis of deployment choices, implementation challenges, and strategic imperatives.

2.1 Introduction

The challenge of spectral optimization continues to constrain mobile networks, particularly under the unfulfilled potential of MU-MIMO. Despite being standardized, MU-MIMO's complexity has hindered wide-scale adoption. Cohere’s USM redefines MU-MIMO control plane execution using Delay-Doppler domain techniques. Yet, its deployment trajectory reflects the intricate vendor-telco dynamics in the modern RAN.

2.2 What is USM?

At the heart of the discussion is the USM, developed by Cohere Technologies, or Cohere. Cohere, headquartered in San Jose, California, specializes in software in the areas of channel detection, estimation, prediction and precoding. Cohere derives its unique selling proposition from its expertise in the Delay Doppler Domain. USM is the primary offering of Cohere – it’s the software that drives multi-user multiple input multiple output (MU-MIMO) Scheduler based on the Delay Doppler domain.

Cohere launched the scheduler in June 2021 in collaboration with VMware RAN Intelligent Controller (RIC).

USM targets MU-MIMO implementation technology. Why MU-MIMO?

MU-MIMO technology enhances cell capacity or spectrum efficiency by facilitating simultaneous communication with multiple users in identical time and frequency slots. MU-MIMO allows concurrent communication with several users in the same slots, thereby increasing the amount of data transmitted per unit of time and frequency and enhancing the efficiency of the most critical resource in the RAN – the spectrum.

MU-MIMO has its own implementation challenges.

The control plane for MU-MIMO requires sophisticated strategies for grouping users for simultaneous communication (addressing the user pairing problem) and for precoding their transmissions in a manner that allows each user to discern their own transmission with minimal interference from others (solving the user precoding problem). Without advanced pairing and precoding, MU-MIMO could potentially degrade performance compared to non-MU-MIMO systems. Despite being standardized by 3GPP, these control plane complexities have limited MU-MIMO's practical application.

MU-MIMO is thus the proverbial holy grail of fine-grained TDD and FDD multiplexing.

What does USM do to address these control plane complexities, and how?

The what is simple to state. USM dynamically establishes orthogonality between paired UEs, reducing inter-beam interference even with minimal angular separation. The establishment of orthogonality is easier said than done – which brings us to the 'how' part.

Traditionally, channel separation is accomplished in the time-frequency domain calculations. USM turns this paradigm on its head and relies on two other parameters – Delay and Doppler.

2.2.1 The Delay-Doppler Domain

Where does the Delay-Doppler domain figure in the USM?

At this time, some context surrounding the phrase 'Delay-Doppler' is required. It is easy to misinterpret this phrase as a special use-case of the Doppler effect – it is not. Delay-Doppler is a geometric representation of the channel composed of delay (distance), doppler (velocity), and amplitude dimensions.

The relevance of the time-domain based Delay is well-understood in the context of waveform analysis.

Why is Doppler relevant in the context of mobile telephony? Consider this – the end-user is mobile or can be mobile. Due to the mobility aspect, Doppler provides a robust framework model for explaining how movement affects the frequency response. Doppler becomes increasingly important as successive mobility generations opt for higher carrier frequencies. High Doppler channels have been found to be extremely effective in eliciting Orthogonal Time Frequency Space (OFTS) modulation performance instead of conventional time-only and frequency-only performance. Delay-Doppler also lends well in situations involving extremely dynamic channel responses. Importantly, Delay-Doppler domain supports separability, thereby addressing the most pressing challenge confronting MU-MIMO.

Cohere postulates that tapping the channel simultaneously for time delay and Doppler effect provides a unique profile of the signal path from the antenna to the user – specifically, the scattering of signal and its causes can be identified with pin-pointed accuracy and near-real time response. Cohere's rationale is in line with multiple studies being conducted by institutions globally.

At this point, let us ask ourselves the obvious question - **What stopped OEMs from using Delay Doppler earlier?**

OEM base stations today use the Time-Frequency channel model. They have never adopted the Delay-Doppler model as it was extracted by Cohere's mathematicians from the OTFS wireless system that Cohere developed as a candidate, but not selected, for 5G air-link technology. What USM has proven is that Delay-Doppler can be used as part of jointly scheduling the base station's downlink queue in an innovative way using the E2 API.

In essence, USM

- leverages existing UE feedback for channel measurement and spatially multiplexes a mix of 5G devices to fully utilize the available spectrum.
- supports massive MU-MIMO deployment capabilities in FDD and TDD under all mobility conditions, a significant advancement given the range and resilience of low-band spectrums.

2.2.2 The AI Approaches

Delay Doppler is not without its challenges. Some of the challenges include finite delays leading to inconsistent power peaking, inter-symbol interferences and imprecise sampling. The most pressing challenge is that of transceiver design that can discern what the Delay Doppler domain is generating. This is where AI comes in.

AI has been pivotal to Cohere's journey. Essentially, the company sees the cellular network as borderless, wherein the transition from one cell to another is envisaged as being seamless without any hard-and-fast model to predict the network behavior. AI is perfectly suited for such a paradigm. In that sense, the Delay Doppler model learns channel behavior on the go making it an ideal foil AI.

Here, the team from Cohere has considered a multitude of possibilities to optimize the channels for better pre-coding, MCS, scheduling, reduced interference, coordinated scheduling, and other factors.

Agentic AI with large language model (LLM) will drive USM in the future. Agentic AI, with its four-step approach to solution formulation is ideal for USM. Let us quickly look at the steps involved and how they make sense for USM

- Perceive: Data gathering from multiple sources and data extraction

- Reason: LLM based solution orchestration by utilizing specific retrieval algorithms for accessing proprietary databases
- Act: Multidimensional execution of the solution thus formulated
- Learn: Consistent and continual improvement by collection of feedback

It is easy to see that Agentic AI is tailormade for USM. USM requires data related to signal strength and phase from a multitude of disparate elements in the signal path. This data is critical to determine the Delay Doppler values. Agentic AI supports this requirement with the 'Perceive' step. The 'Retrieve' step is also of significance, as it supports an intensive solution finding approach concerning proprietary databases. Cellular mobile network data structures and formats fit the above database description perfectly.

2.3 The Dilemma and the Workaround

Cohere is a promising expert in a very specific niche of the RAN functional chain – channel estimation. The USM has demonstrated quantum improvements in channel lifetime – up to an impressive 50%, providing an effective and quick workaround to the pressing issue of spectral usage maximization. With such commanding numbers, Cohere should be having the ultimate say in the USM deployment methodology. In terms of a forward-looking approach, the xApp is clearly the front-runner. This would have been true in a market with a more democratic primary customer landscape and a more proactive secondary customer landscape. That is clearly not the case in the present-day cellular mobile market.

Cohere seems to be finding its delivery choices being dictated by its customer profile. Cohere's primary customers are the RAN OEMs while its secondary customers are the cellular operators, or telcos. Cohere deals with both these customers directly, without intermediary value-adding entity.

The present-day cellular mobile RAN market is a juxtaposition of contrasting trends.

On the technical side, the industry is an overdrive exploring newer horizons in the areas of AI adoption, opening of the RAN interfaces, network slicing and facilitating the movement towards eventual adoption of 6G technologies by the end of the decade.

The picture however, is starkly different when the state of the vendors is assessed. The vendor landscape is remarkably frozen in its composition, with the number of key vendors progressively shrinking.

There are about 1000 cellular operators present globally. Their higher numbers however do not guarantee any leverage for Cohere in terms of choice of delivery. Most of the telcos are struggling to recover their returns over investments made in 4G and 5G networks. The appetite for trying out radically new delivery approaches like the xApp remains restricted to very large tier 1 telcos.

The consequence of this dichotomy is that every seminal technological development emerging from any quarter has to ultimately acquire a buy-in from either the larger RAN vendors or from influential cellular operators. Each of the above entities have their peculiar pathways and reservations for USM deployment. Cohere has had to therefore navigate its deployment strategy deftly, addressing the concerns of the OEMs and the telcos.

As it turns out, xApp is not the preferred delivery option for Cohere.

At this stage, we know the following:

- The channel allocation mechanism has been traditionally based on time-domain parameters
- Cohere suggests an alternative methodology – the Delay Doppler. This methodology has demonstrated performance improvement up to 50%.
- This methodology drives the USM, the brain behind the channel scheduling mechanism.

In any other market, a technology offering such impressive improvement would result in potential customers facilitating its internalization. The cellular market is different though -given the ever-shrinking number of consequential OEMs. Consequently, Cohere needs multiple vehicles to reach the heart of this market.

Let us consider the approaches one-by-one in subsequent sections.

3. THE XAPP

This section examines the xApp-based delivery approach for the USM.

To make USM more appealing to telcos, Cohere's approach involves decoupling RAN intelligence (specifically for MU-MIMO scheduling, including dynamic user pairing and precoding) from the Distributed Unit (DU), running it as a cloud service atop near-Real Time RIC, to start with.

Decoupling the RAN intelligence is a bold step. Traditionally, schedulers were deemed too latency-sensitive to be moved out of the DU. RAN demands signalling latency to be in milliseconds. Latency is a function of distance. DU provides adequate proximity to keep latency under control.

While hosting the scheduling function in the DU addresses the latency challenge, it falls short of providing a holistic command and control framework for MU-MIMO performance. This objective can be met only with some level of centralization. Herein, Cohere has leveraged the graded application hosting architecture offered by the Open RAN framework.

The O-RAN architecture disaggregates the RAN into the following components and network functions:

- Near-Real-Time RAN Intelligent Controller (Near-RT RIC): This component provides real-time control and optimization of RAN elements and resources using fine-grained data analytics. It uses a set of standardized "xApps" which can be developed by third parties to optimize network behavior.
- Non-Real-Time RAN Intelligent Controller (Non-RT RIC): This is responsible for non-real-time aspects of RAN optimization, policy management, and providing a platform for AI/ML-driven insights and optimization. It uses a set of standardized "rApps" which can be developed by third parties to optimize network behavior.

Cohere offers its MU-MIMO scheduler as xApp.

It is not difficult to see why. rApp are suitable for control loops with response of 1 second and above. This response time is eminently unsuitable for the low-latency demanding MU-MIMO scheduler. xApp, on the other hand, are more ambitious with control loop responses in the range of milliseconds.

3.1 THE APPEALING FACTORS

There is considerable merit in decoupling the scheduling function and moving it closer to the core. Cohere, in fact, claims that this decoupling allows for more effective channel estimation and prediction, enabling MU-MIMO's application to existing LTE handsets and increasing spectral efficiency in LTE networks without additional spectrum usage or handset upgrades. These benefits also extend to 5G deployments.

Let us now discuss certain specific advantages of the xApp approach.

3.1.1 Direct Interface to the Base Station

An under-appreciated aspect of the Open RAN ecosystem is that it offered a direct interface to the base station. Cohere acknowledges that this is something that could not have been achieved with any of the OEMs. Open RAN provided them with the platform to showcase the USM as a credible solution. The successful RIC-based trials have undoubtedly played a pivotal role in opening doors for USM at telcos and OEMs alike. Indeed, it was the Open RAN that brought Cohere together with Vodafone, VMware, Capgemini Engineering, Intel and TIP way back in 2021.

It would be reasonable to surmise that RIC provided the ultimate PoC for the USM.

3.1.2 Minimal Additional Overheads

There is apprehension that the insertion of another entity (RIC) in the channel scheduling mechanism may lead to the performance being compromised. On this account, Cohere reports that there is no deterioration in the IO rates when USM is hosted as xApp. Cohere also affirms that the CPU usage of the xApp is also range-bound and predictable.

3.1.3 Latency

RIC presents an additional hop in the network. If hosted in the cloud, that hop may traverse public networks. Latency values become consequential in such scenarios. Cohere caps the acceptable latency for USM at 50 milliseconds. This requirement can be satisfied in most mid-sized countries with smart placement of data centers. In smaller countries such as in the EU, the end-to-end latency rarely exceeds 10 milliseconds. The challenge that Cohere can face is networks with larger geographical spreads with lesser population density. But then, it is important to remember that the lesser the population density, the lesser the load is on the spectrum. The USM was designed to address the challenge of channel allocation in regions with high user-density. Such regions also witness a high density of data centers, as it is economically viable to do so.

In summary, the challenge associated with latency may not be as grave as it may initially appear.

3.1.4 Costing Comparison

On a one-on-one basis, a single installation of USM xApp interfacing with a single telco base station may appear to be a costlier proposition than the software being loaded directly on the base station of the same telco as there is the additional element of RIC coming into play. This comparison is however misleading due to the following reasons:

1. Single instance of xApp can interface with multiple base stations of a single operator simultaneously whereas the in case of direct to base station, the interface with each individual base station has to be established separately.
2. A single RIC interfaces with multiple xApps which can have overlapping mandates about the data sought from the base stations. The cost of setting up dedicated interfaces for each of these xApps individually can be prohibitive.

In the final reckoning however, these disparities in costing are a small fraction of the overall running expense of a base station. By all estimates, electricity contributes more than 70 percent of the running expense of the RAN. Ericsson, in fact, pegs the percentage between 80 and 85 percent. In the post covid era, global electricity costs have risen sharply. In February 2025, the International Energy Agency reported that the average wholesale prices for EU in Q1 of 2025 was more than 60% higher than the corresponding figure in the Q1 of 2019. It is noteworthy that this gap persists even after the nearly 60% drop from the maddening peaks of Q3 of 2022 in the aftermath of the Ukraine war.

In essence, operating costs of the USM xApp remain a tiny fraction of the overall RAN costs. Even otherwise, USM xApp is bound to deliver lower operating costs as the scale of adoption increases.

3.1.5 Scalability

In case of the USM xApp, scalability boils down to the following factors:

- Monitoring busy sectors
- Tracking and measuring IO activities in these sectors
- Tracking the state information of the mobile devices

The USM xApp housekeeping and maintenance has to essentially ensure that the requisite resources such as CPU processing cores and RAM quantity keeps pace with these demands placed by the above mentioned factors.

How ready is the RIC to manage these demands?

Here is the answer. RIC multiplexes multiple requests to the base station. If there are multiple xApps wanting the same information from the base station, the RIC makes a single request to the base station and hands back the information to the requesting xApps

They way RIC is envisioned, scalability is a given. RIC is fault-tolerant and its framework is based on distributed databases. As it is cloud-native and thus driven by microservices, scalability can be achieved at a granular level.

3.1.6 Security

USM operates at the very heart of the call management methodology. Any compromise to its integrity can adversely affect channel management and can all threaten the integrity of the traffic itself.

There are bound to be apprehensions about the security aspects of USM as xApp on RIC, a resource that can be hosted in the cloud. The multidimensional nature of the RIC – input aggregator, output distributor and application interface manager with xApps and rApps from multiple vendors, can also add to the uncertainty when compared against the tried and tested predictability of the direct-to-base station approach. The security challenges are amplified in 5G NR environment as its dynamic nature requires continual authentication of every device and applications. This is in addition to preventing intrusions and ensuring end-to-end integrity of the network.

The above challenges are well-understood.

RIC offers end-to-end encryption. The O-RAN Security Requirements Specifications v2.0 specifies initial security requirements per interface and per component. Features include confidentiality, integrity, and availability (CIA) protection for all interfaces; TLS 1.2+ and Public Key Infrastructure X. 509 (PKIX) for Fronthaul M-Plane mutual authentication; securing the software bill of materials for every delivery in accordance with NTIA guidance. Some of the other key specifications include O-RAN Security Protocols Specifications v3.0 and O-RAN Security Tests Specifications v1.0.

Suggestions are forthcoming from industry as well as governments. In its 2022 report about Open Radio Access Security Considerations, CISA outlined the following approaches for xApp and rApp security - Usage of secure AI and ML data sets and models; Secure peering between inter xApp, non-RT to near-RT, and xApp to near RT using mutual authentication; Conflict mitigation strategies for parameters; and Multi-factor authentication for data sources, among others.

All-in-all the security aspect surrounding the xApp based approach is well understood and appropriately addressed for telcos willing to invest resources meaningfully.

3.1.7 Dynamism in Management

USM as xApp lends dynamism to the management at levels very difficult to attain using the direct to base station approach. The reasons are not very hard to seek. Open RAN is crafted for a multi-vendor, multiplatform environment. RIC is based on microservice architecture, which lends itself to granularity, making it possible to tweak parameters with pin-point precision. The RIC framework takes into account the possibility of distributed data sources and provides secure and speedy means of accessing and leveraging these resources. This centralized approach is extremely important to leverage the benefits offered by AI-based analytics more effectively.

3.2 Customer Response

The merits discussed above are not restricted to the theoretical domain. Cohere has successfully demonstrated, during its trials with Vodafone Spain in 2024, that its approach is able to deliver savings up to 50%. The trial involved the usage of USM in the xApp mode in a proper Open RAN setup. Vodafone and Cohere tested USM in the 2.1 GHz band in 10 MHz spectrum. The trial covered indoor and outdoor settings. Cohere's engagement with Vodafone and Open RAN is of a much older vintage. Way back in June 2021, the two companies along with VMware, Capgemini Engineering, Intel and TIP demonstrated 5G MU-MIMO run successfully on a multivendor Open RAN RIC.

Cohere's xApp gambit is not without its supporters. RIC vendors are among the strongest backers. Cohere counts VMware, Juniper and Capgemini among its vendors. Mavenir is a major collaborator as well, having integrated the xApp in its RIC. The prospect of cracking the L1 disaggregation challenge would have understandably enthused the RIC community.

3.3 The Business Challenge

The above factors notwithstanding, xApp is not the most acceptable delivery option for Cohere's customers. Why should that be the case?

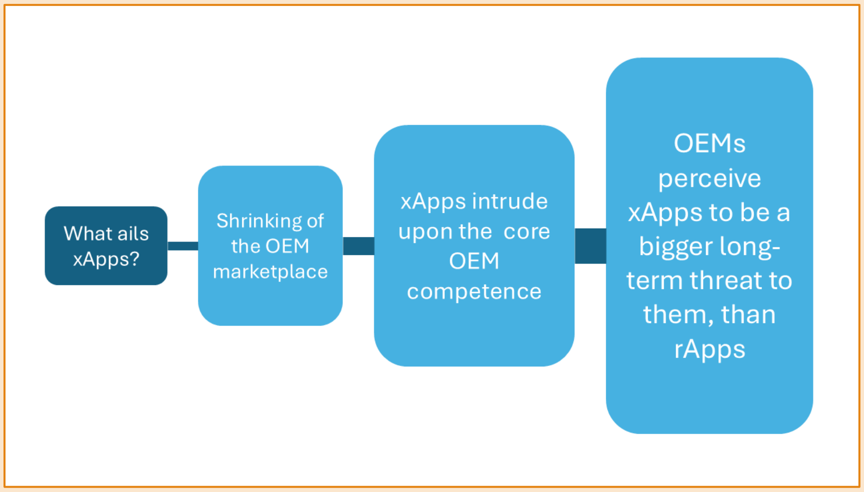
The real problem is clearly on the business end.

What happens next after the successful demonstration of 50% spectral gains? The stakeholders expect vendors to incorporate open interfaces that can support USM in their interfaces. Will that happen? Yes, in an ideal world, as the capacity improvement can be accomplished without changing the equipment on either the user side or the telco side such as radios and antennas. The only requirement is support for Open RAN, and by extension, the RIC.

Let us look at the RIC now. The RIC by itself heralds the opening of the RAN to third-party developers. Consider it as the RAN equivalent of Apple Store or Google Marketplace. In an ideal world, vendors should be happy, knowing that equipment enhancements or upgrades can be completed with minimal service disruption. The current state of the RAN equipment market puts paid to hopes of vendors embracing USM via RIC.

Figure below summarizes the business reasons slowing down the march of the xApp.

Figure 3-1: What ails xApps


What ails xApps?
Shrinking of the OEM marketplace
xApps intrude upon the core OEM competence
OEMs perceive xApps to be a bigger long-term threat to them, than rApps

Source: Insight Research

The RAN equipment vendor market has presently shrunk enormously due to two factors – progressive acquisitions of competitors by established players such as Ericsson and Nokia; and an embargo of sorts imposed on Chinese vendors by the West and vice-versa. This situation bestows unfair advantage on the OEMs that have little incentive in allowing serious value addition by third parties. Even within the RIC, it is the Near-RT RIC that poses the more pressing threat to the pre-eminence of the OEMs. There is an evident lack of enthusiasm and support for rApps from established vendors such as Ericsson. Their argument is simple – why opt for xApp from another vendor when the same functionality is available in the RAN itself?

It can be rightfully surmised that the entry barrier is steeper for xApps than for rApps. This could be because rApps, being closer to the SMO in the Open RAN, may actually help off-load functionalities that the mainstream OEMs consider peripheral; to other vendors. In other words, OEMs perceive rApps to be less threatening than xApps.

This threat perception surrounding xApps in the OEM mind space is the single biggest impediment for the xApp based approach for USM.

4. Direct-to-Base Station

This section examines the direct-to-base station based delivery approach for the USM.

In response to the vendor reluctance to xApps, Cohere transitioned into plan B, which is overlaying the SM on existing vendor hardware. This approach maintains the essential structure of the RAN unaffected. The RAN structure preservation makes this approach appealing to a large number of operators who would want to experience the benefits of USM without getting into the Open RAN mode. This is ironical by itself, as Open RAN was piloted by operators themselves. Due to vendor consolidation, Open RAN and the concomitant RIC framework is less appealing than it should be. Most importantly, the direct-to-base station approach opens doors for the USM to work with RAN gear supplied by a diversity of vendors.

Cohere decided to load the USM on its own hardware and then interface it with the RAN. The deployment modality is thus very straightforward – the USM can be loaded on a central server or it can be collocated with the gNodeB.

4.1 What works?

Direct-to-base approach has some very obvious limitations vis-à-vis xApps. It needs greater customization and it is more resource intensive. It is however, not bereft of benefits. This section covers some of the obvious and not-so-obvious strengths of the direct-to-base station approach.

4.1.1 Elegance of gRPC

Cohere uses the exact same APIs regardless of the approach. In the direct-to-base-station approach, Cohere utilizes gRPC, a remote procedure call (RPC) framework developed by Google. gRPC is presently open source. gRPC provides an elegant pathway for the integration of USM with the base station.

Let us look at gRPC more closely.

RPC by itself is one of the oldest and most widely employed API building methodologies. Being the oldest, it is suited for more straightforward implementations. It works best in situations where there not too much of a difference between calls to the local system and calls to a remote system – because that is exactly how the procedure calls in RPC work. RPCs are therefore not cut out for complexity of messages and diversity of systems. They are vulnerable to exposing internal implementation methodologies. Alternative approaches such as representational state transfer (REST) attempted to address some of the limitations such as interoperability. What REST did was very intuitive. It embedded all the identification within the API itself. It is fair to surmise that RPC was more trusting of the target interface procedures and therefore more vulnerable to altering its own performance due to the vagaries of the procedures; while REST tried to inject consistency in its performance by overburdening its messages with heavy metadata.

So why gRPC, and why would Cohere opt for that?

gRPC attempts to retains the advantages of RPC while addressing a whole lot of its limitations. Thus gRPC definitely scores over REST in terms of message size and the raw performance of API itself. While doing so, it elevates the game of its predecessor RPC by offering more connection options and in-built code generation. These improvements are results of employing HTTP/2 and protocol buffers. HTTP/2 is an advanced version of HTTP/1.1 employed by RPC and brings to the table, binary messaging instead of text and parallel processing of request instead of serial. Protocol buffering scores over the XML and JSON methodologies in REST in terms of CPU utilization and preserving message integrity.

It is important to note that gRPC suffers from significant challenges such as lack of browser support and lack of user maturity.

So why would Cohere opt for gRPC? We can derive some plausible reasons.

- gRPC does not compromise speed
- gRPC has fewer messaging overheads
- gRPC is more versatile
- gRPC utilizes CPU resources with greater efficiency
- gRPC preserves message integrity

All of the above attributes make gRPC a suitable wrapper of the USM API, which remains unchanged, regardless of the deployment option.

4.1.2 Acceptance in Low and Mid-band FDD

MU-MIMO has found ready acceptance in the higher reaches of the spectrum, where TDD is the dominant technology. However, a great majority of the base stations remains firmly in the low and mid-band FDD technology driven spectrum. These base stations are not going anywhere and they are not compatible with RIC paradigm. Why is this incompatibility? The reason is simple – telcos will need to invest upgrading these base stations to adopting RIC and E2 interfaces. This is a costly proposition. The sheer number of these low and mid-band base stations – estimated to be in the range of 10 million (of which, 4 million are outside China – making it a direct addressable market for Cohere. Why should Cohere let go of this captive market, which will be ready to jump over to achieving greater spectral efficiency at very marginal incremental investment? That too, when there is a clear pathway enwrapped in the versatile gRPC?

It makes eminent business sense for Cohere to therefore keep its doors open for legacy low and mid-band FDD-driven base stations using the direct-to-base-station route.

4.1.3 Eschewing the Cloud and its Vulnerabilities

In Cohere's experience, barring exceptions, telcos have clearly not reached the confidence to outsource the hosting something as critical such as network scheduler to something as vulnerable as the cloud.

One of the major drawing point of the RIC is its cloud-based approach. Having a bouquet of network configuration xApps for the telco to deploy is the near perfect freedom of choice. The cloud is not without its risks. One can say that GCP, Azure and AWS is the leading troika of cloud platforms. Each one of them has exhibited vulnerabilities- some graver than the other.

In July 2024, Microsoft 365 and Azure witnessed a global outage of close to eight hours. This outage was triggered by DDoS attack attributed to hacktivist groups. The concerning fact was Microsoft's defense system – the Azure DDoS Protection Standard ended up aggravating the situation due to faulty implementation affecting the performance of its CDN, among others. The outage had wide-ranging impact on multiple business sectors including travel, finance and others. Azure continues to be deployed heavily by telcos for domains spanning from network functions to network management. Azure itself lists AT&T, Lumen, Orange, Telkomsel and Vodafone among its key customers.

Let us be clear here. The purpose is not to single out Azure from its peers. Other platforms have exhibited similar vulnerabilities leading to varying levels of disruption.

What do telcos think about the cloud challenges?

The Infosys Cobalt Survey covering telecom operators published in October 2024 highlighted the following observations:

- More than 90% of the respondents employed more than 1 CSP, clearly highlighting the importance of risk mitigation
- Less than 50% of telcos have actually consumed the cloud services that they have committed to, highlighting apprehensions in moving legacy, mostly time sensitive operations to the cloud
- Monitoring cloud costs was a significant bottleneck identified by more than 50% of the respondents

Channel scheduling has to be operational 24x7. As USM aims at improving channel lifespan, its response has to be dynamic and real time, leaving little tolerance for latency. The cloud experience thus far indicates that its migration to the RIC has to be qualified with appropriate failsafe measures and performance guardrails.

4.1.4 Avoiding RIC Management Overheads

RIC provides freedom of choice. Choice, by its very definition, leads to diversity. Managing diversity is complex and resource intensive. Here, the RIC is at a disadvantage. Being a shared resource, telcos have to device better grip on the RIC resource allocation itself – the two-way communication, for instance, between the RIC and the NE should not be compromised due to uncertainty in the availability of RIC resources. Direct to BTS deployment on the other hand, does not suffer from such uncertainty. Here, the telco, is more surefooted about the USM getting the access to the requisite resources.

Continuing with this thought, for a good amount of telcos, the RIC itself is a significant management overhead. And here we are including tier 1 telcos as well. What should telcos focus on ideally – improvement in channel age due to USM performance, or on the nitty-gritties of managing RIC?

RIC entails a spectacular increase in the complexity of APIs, and we are not just talking of the USM API here. RIC also places demands of interoperability among applications, platforms and modules. If not managed well, RICs are vulnerable to conflicts. In fact, the O-RAN Alliance itself has identified and categorized near-RT conflicts as direct, indirect and implicit. Direct conflicts pertain to the same parameters being accessed controlled by multiple xApps. Indirect conflicts pertain to related (not the same) parameters being controlled by different xApps. Implicit conflicts pertain to those where the affected parameter may not have an obvious direct relationship.

In a whitepaper published by i14yLAB in October 2024, the authors belonging to Nokia, Vodafone, Fraunhoffer HHI, Rimedo Labs, Technische Universitat Berlin and Capgemini, postulated the following reasons for conflicts in the Open RAN architecture:

- Gaps in addressing complex interactions between multiple intelligent Apps
- Conflicting objectives of the Apps
- Interoperability challenges

The same whitepaper also discussed the possibility of horizontal (rApp-rApp and xApp-xApp) and vertical (rApp-xApp) conflicts.

Conflicts require elaborate mitigation strategies involving elaborate KPI management and complex simulation tools.

The question to be asked at the end of the above discussion is this - What would the average telco want to focus on in the short to medium term – immediate performance improvement sans complexity as promised by the direct-to-base state approach; or RIC driven approach where gains are distant but complexity is immediate?

4.1.5 Eliminating Additional Stakeholders

RIC has its advantages and its challenges. The net effect of these contradicting impulses depends on the context. This context is defined by the ability of the telco to handle complexity. Shorn of all these impulses, the fact remains that RIC is another component that needs to be integrated solidly to make the USM work as per plan. Integrating every additional component places extra demands for having the proof of concept (PoC) ready. The rate of failure due to inadequate PoC is significant – especially when the technology in question is heavily reliant on AI. PoC has to be taken seriously therefore and that entails the following overheads:

- PoC definition and scoping
- PoC design and tracking
- Testing and documentation
- Project management

In contrast, the direct-to-base station approach can sidestep all these requirements and expedite deployment.

Expeditious deployment was important in case of Bell Canada, which offered a very short time frame to Cohere to make the integration work. In that episode, Bell Canada's legacy base station base did not support RIC natively due to the lack of E2 interface. In such situation, the direct-to-base station approach was very effective.

4.1.6 API Emulation

A reasonable advantage that the RIC brings to the table is that once the xApp is established on the RIC, unit testing for individual OEMs is not required. RIC thus provides blanket compatibility across the board.

Direct-to-base station approach, on the other hand, requires Cohere to unit test its USM with every single OEM base station model. This process can be cumbersome.

Cohere has addressed this issue by developing an emulator for the USM API.

The emulator facilitates the software developers at the OEMs to test each and every call on the API to check the quality of the data and then respond with the appropriate information. The emulator is a major confidence booster for OEMs. It ensures that internal operations of the actual USM machine didn't get in the way of unit testing. The exhaustive unit testing also ensures an almost plug-n-play interfacing of the USM with the base station.

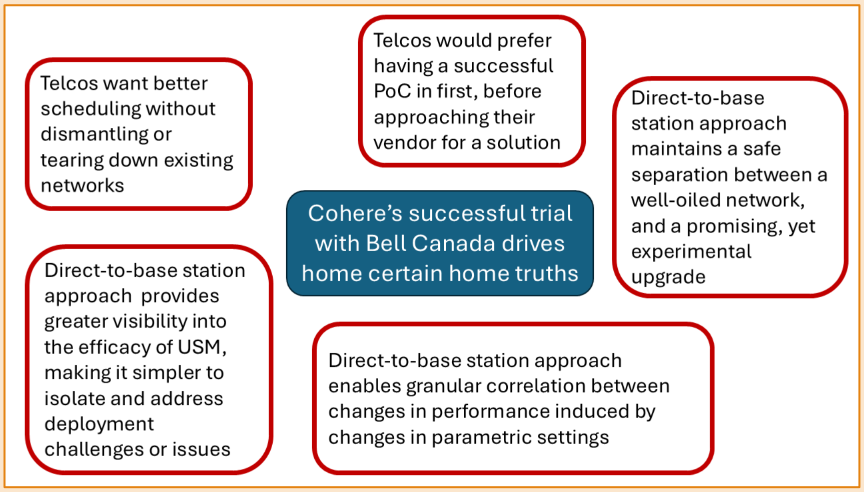
Cohere thus smoothens the rigor involved in the direct-to-base station approach by its solid emulation process.

4.2 Customer Response and Inferences

For the so-called brownfield networks therefore, Cohere had to opt for dedicated hardware. Cohere successfully demonstrated this approach with Bell Canada in early 2025. Here, the trial involved the 850 MHz FDD band in a 5G SA setting. The choice of the band is pertinent, as it demonstrates that enhanced MU-MIMO scheduling is possible even at lower frequencies – something that was thought of being confined only to the upper reaches of the frequency bands. Of equal salience is the FDD band, which was hitherto considered to not be amenable to MU-MIMO.

Figure below drives home some inescapable conclusions from the Bell Canada experience.

Figure 4-1: The Direct-to-base station appeal


Telcos want better scheduling without dismantling or tearing down existing networks
Telcos would prefer having a successful PoC in first, before approaching their vendor for a solution
Direct-to-base station approach maintains a safe separation between a well-oiled network, and a promising, yet experimental upgrade
Cohere's successful trial with Bell Canada drives home certain home truths
Direct-to-base station approach provides greater visibility into the efficacy of USM, making it simpler to isolate and address deployment challenges or issues
Direct-to-base station approach enables granular correlation between changes in performance induced by changes in parametric settings

Source: Insight Research

In conclusion, there are obvious and less-than obvious advantages of the direct-to-base station approach. This approach provides an effective pathway for telcos and OEMs eager to capitalize on the superior channel scheduling methodology of USM expeditiously, without risking entering into a completely new Open RAN paradigm.

5. Joint Scheduling

This section examines the joint scheduling-based delivery approach for the USM.

The xApp and the direct-to-base station approaches essentially feature Cohere playing catch-up with existing RAN products. In both cases, the USM has to interface with existing RAN stack.

Considering the handsome gains USM has demonstrated in trials, it is not inconceivable that OEMs will want to embed the USM logic onto the device firmware itself. Is that a possibility that Cohere has budgeted for?

The answer is yes. Cohere has developed what it calls a technique called joint scheduling. This section provides an overview of the joint scheduling approach and the significance of dApps in the execution of this approach.

5.1 Modus Operandi

Here is how joint scheduling works:

- The USM sits between the base station and scheduler. It can be integrated into the DU itself. The USM continually listens to the information about the scheduling load from the base station.
- Based on the inputs received, it returns to the scheduler with its scheduling recommendations.
- The exact positioning of the USM in the DU can follow multiple approaches. It could be integrated into the MAC layer of the DU; or it could be hosted as a distributed application, or dApp.

Integration of USM into the MAC layer is the ultimate in embedding it with the OEM hardware. This approach will require Cohere to surrender nearly all control of the functioning of the USM to its OEM customers.

How does Cohere then maintain a modicum of control in the OEM domain? Enter dApps.

5.1.1 dApps and Their Role

It is important to dwell into the possibilities of what dApp entails. While xApps and rApps are well understood and extensively discussed; the concept of dApps has not received commensurate attention. For one, dApps are not governed by the RIC. They sit in the DU (or CU) and receive key performance indicators (KPI) from the MAC or PHY layers.

dApps are suitable for operations with time-scale in the sub-10 millisecond range; making them faster than rApps. This is not surprising, given that they are closer to the subscriber than rApps. Certain dApps that target downlink and uplink management of the end-devices are designed to operate in the sub-1 millisecond range, which make them the truest real time among all the apps discussed. USM may be less aggressive as it targets the scheduling function involving multiple end-devices concurrently. Their faster response times make them suitable for mining rich AI data that xApps and rApps cannot lay hands on.

dApps offer an elegant path for Cohere to maintain architectural autonomy over the USM even while working closely with the OEMs at the architectural level.

Regardless of the joint scheduling mechanism being executed using the dApp or embedding the USM logic into the MAC layer, one thing is clear - the USM does not take over the scheduling function; it merely supports the scheduler with qualitatively better channel management inputs.

5.2 Customer response and Inferences

So we know that it is possible to overlay USM on existing BTS/gNodeB as software or patch. Has this gambit succeeded so far? Unsurprisingly, no.

Few vendors (and their customers) are amenable to this approach. The hesitation of vendors is understandable to an extent – Cohere is possibly highlighting a gap in their product design. What about the telcos though? What explains their reluctance? The answer is equally straightforward – vendors and telcos and locked in a tight embrace which curtails the telco appetite to experiment with vendor gear without the vendor being on board. It will require considerable boldness on the part of the vendor to accept a teardown of their gear. Thus the conservative approach of the vendor-telco duo constricts deployment options for Cohere.

In conclusion, Cohere faces considerable challenges in persuading its target market in embedding the USM logic into the very heart of their gear.

6. Delivery Approaches Compared

This section presents the conclusions of the previous section by summarizing the multidimensional comparisons of the three approaches.

Table below presents a snapshot of what we know so far about the three approaches.

Table 6-1: Feature comparison of USM deployment modes

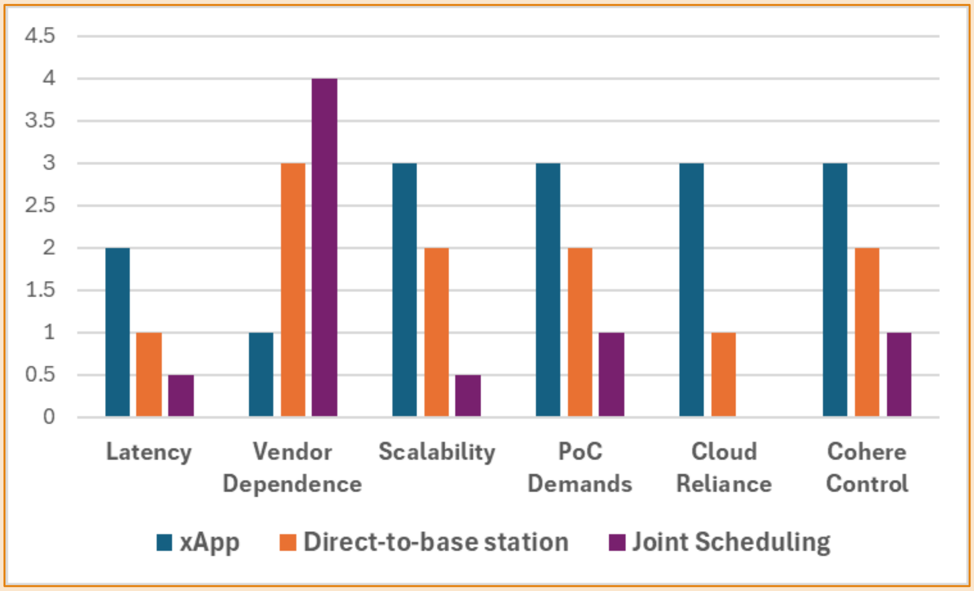
Feature	xApp (RIC-based)	Direct-to-Base Station	Joint Scheduling
Integration Point	Near-RT RIC (via xApp)	gNodeB (via gRPC)	DU MAC Layer / dApp
Latency Sensitivity	Medium (<50ms acceptable)	Low	Very Low (sub-10ms)
Vendor Dependence	Low (vendor-neutral)	High (OEM-specific tuning)	Very High (deep OEM integration)
Scalability	High (via RIC microservices)	Moderate	Low to Medium
PoC Requirement	High (multi-party validation)	Moderate (custom per OEM)	Very High (firmware-level)
Cloud Dependence	High	Low	None
Control Retention (by Cohere)	Moderate	High	Low (control ceded to OEM)
Adoption Scenario	Tier-1 telcos; Open RAN environments	Brownfield networks; low-band FDD	OEM-led embedded evolution
Hardware Dependency	RIC-compatible base stations	Cohere hardware overlay	Embedded in base station firmware
Trial Example	Vodafone Spain, 2024	Bell Canada, 2025	In development / exploratory

Source: Insight Research

How does the above comparison bode for Cohere and its customers in terms of managing deployment complexity?

Figure below encapsulates the complexity indices of various approaches.

Figure 6-1: Complexity* comparison of USM deployment modes



Source: Insight Research

The higher the value, the greater the complexity

Let us understand the implications of the above ratings. Before doing so, it is important to reiterate that the joint scheduling option, with all its advantages, remains a non-starter due to the unwillingness of the OEMs to on-board the USM logic on their products.

- Latency: It is obvious that joint scheduling will deliver the best latency performance as compared to the other two modes. The differential, however stark, is of little consequence in the final reckoning as the USM is capable of tolerating latency up to 50 milliseconds, something that is comfortably achieved by the xApp deployment.
- Vendor dependence: xApp guarantees near-total vendor independence, while joint scheduling binds Cohere tightly to the vendor stack. The deployment of gPRC wrapped API in the direct-to-base station approach provides the perfect balance of neutrality from Cohere's perspective and ease-of-adoption from the vendor/telco perspective.
- Scalability: It is important to realize that while xApp is largely synonymous with scalability, the direct-to-base station approach is not very far behind either. The API remains the same in all deployments. Cohere has also developed an effective emulator, reducing the customization burden

considerably. Consequently, the direct-to-base station approach is not exactly lacking in scalability either.

- PoC demands: This comparison takes into account an apple-to-apple scenario. Thus it compares a single deployment for a single base station. In such cases, it is easy to understand why the burden is higher for the xApp. There is an addition RIC to reckon. This burden evens out as the number of deployments increases. Telcos however, are hesitant to graduate into the RIC regime unless their tried and tested OEM vendors are on board. The balance thus tilts in favor of more conventional modes, chiefly the direct-to-base station approach. While joint scheduling offers the least cumbersome PoC pathway, the on-boarding of OEMs on this option is still work-in-progress.
- Cloud reliance: This is a critical point given the repeated expose of vulnerabilities of the cloud makes news. Here again, the direct-to-base station approach is on a firmer footing.
- Cohere control: Cohere exercises maximum control in the xApp option. It has to cede space to its customers in the direct-to-base station approach. This is something that the customer will gladly accept, given that it is bundled with the promise of deployment readiness in the present avatar of their base stations.

Cohere’s strategic flexibility—supporting all three approaches—is likely to serve it well in a surprisingly conservative market which is increasingly under pressure to deliver higher performance without wholesale infrastructure upgrades.