Unleashing the potential of AI in the RAN

Table of contents

Bringing intelligence to the RAN / 03

How AI and machine learning are used in RANs / 05

Understanding the intelligent RAN architecture / 07

Demonstrations of AI in the RAN / 11

- Saving energy through dynamic power management / 12
- Improving spectrum efficiency / 14
- Improving MU-MIMO efficiency / 16
- Improving resource utilization of O-DU upper PHY software / 18
- Improving traffic steering / 20

Accelerating AI on general-purpose processors / 23

Conclusion / 25

Bringing intelligence to the RAN

Mobile data consumption is increasing. For example, GSMA predicts that mobile data traffic in Europe will almost triple between 2023 and 2028[1], while CTIA reports that 2022 saw the greatest increase in mobile data traffic ever in the US.[2] It increased from 53.4T MB to 73.7T MB, nearly double the previous year's increase.

The industry introduced 5G in response to escalating data demand and to enable new use cases. With this new standard came more frequency bands, more radio types, more base stations, and a wider variety of connected devices, resulting in more network complexity. In addition, vendor complexity is growing as the network becomes more open. Network operators need novel ways to address these challenges.

Artificial intelligence (AI) will play a major role in new telco use cases and in optimizing 5G networks. Some operators are already using it to optimize cell site design. The next frontier for AI in the network is optimizing resources and user experience in real-time using data about network conditions and user behavior. In this way, AI can help mitigate network complexity, reduce operating costs, and improve Quality of Experience (QoE). General-purpose processors have evolved to help make AI performant and cost-effective for the RAN.

In this paper, we'll introduce several use cases for AI in the radio access network (RAN):

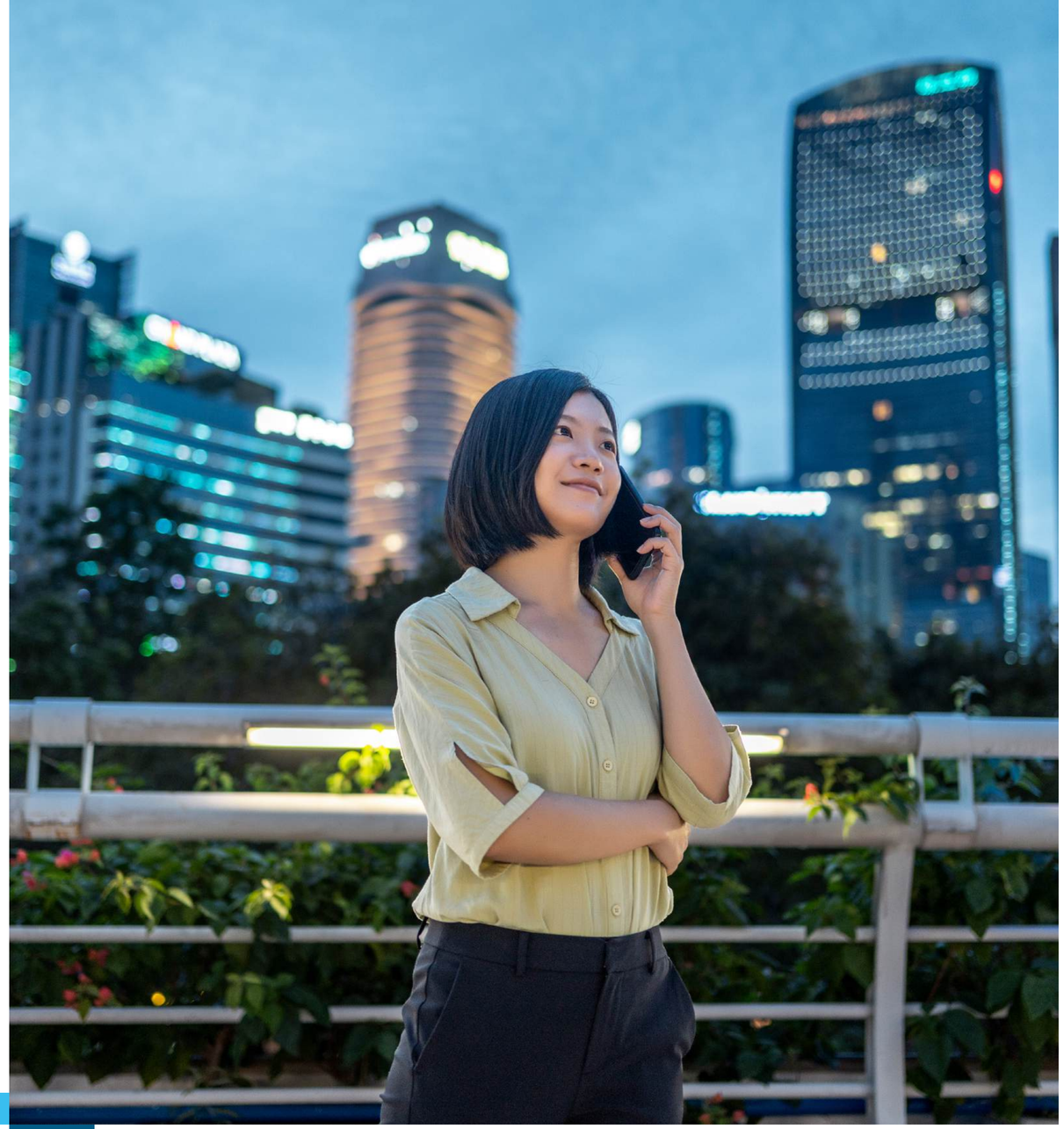
- Saving energy
- Improving spectrum efficiency
- Improving multiple-input multiple-output (MU-MIMO) efficiency
- Improving resource utilization of O-RAN Distributed Unit (O-DU) upper PHY
- Improving traffic steering

We'll highlight demonstrations of these benefits pioneered by innovative virtualized RAN (vRAN) vendors using Intel® technologies.

How AI and machine learning are used in RANs

AI/machine learning (ML) improves network efficiency, performance, resiliency, and reliability and enables new business models. For example, network behavior can be adjusted dynamically to respond to changing network demands.

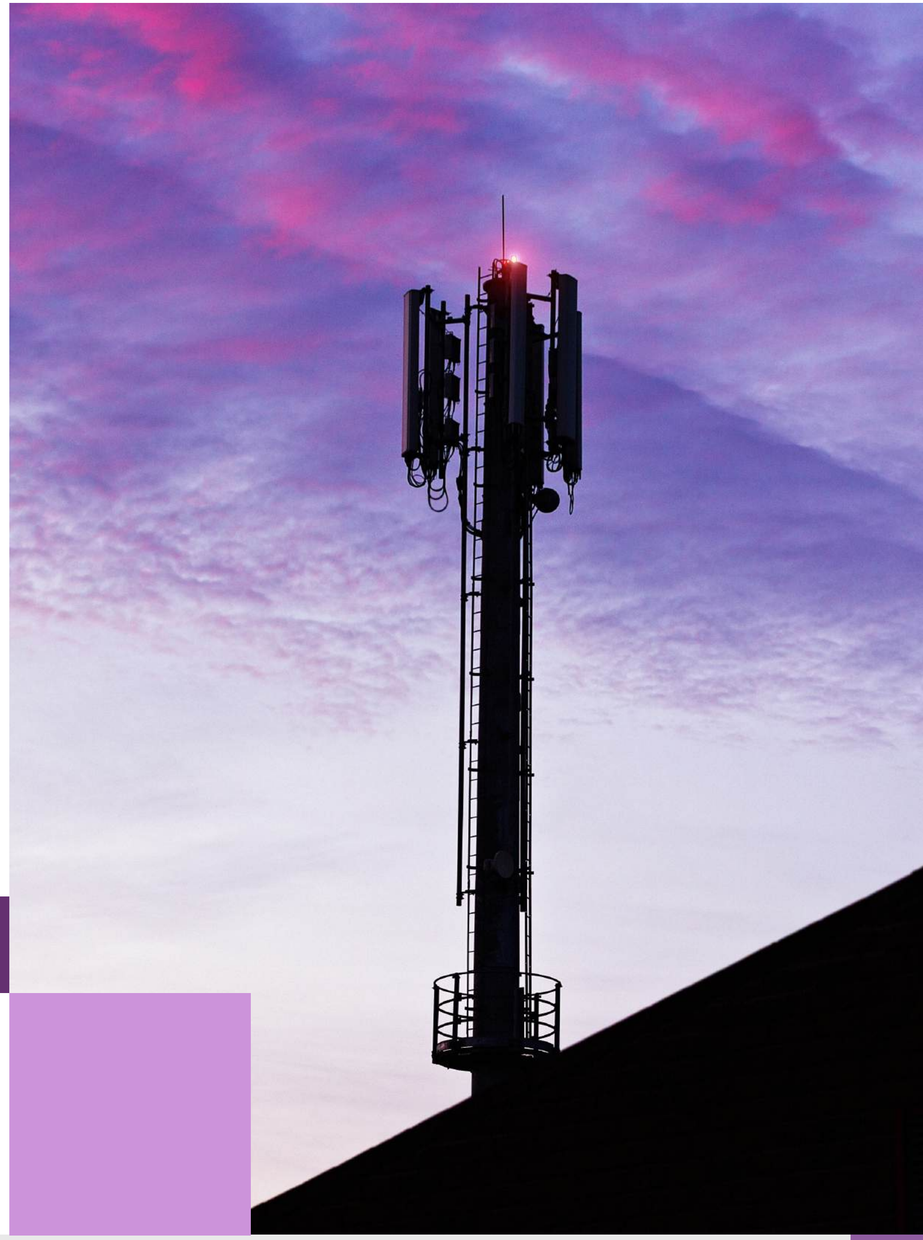
Imagine a situation where user demand occasionally skyrockets in a specific location, such as at a large concert venue. AI algorithms could be configured to predict capacity needs and proactively adjust network resources for increased bandwidth in the venue during a concert. When the crowds leave, network resources are reduced in that area or reallocated to another high-demand area.



These capabilities are powered by new AI prediction, correlation, and recommendation functions to implement a fully automated network. Predictive analytics models use data and trends over time for various use cases, for example, to preemptively optimize network resources and minimize idle or underutilized resources.

These algorithms can be implemented directly in the network's O-RAN Distributed Unit (O-DU), O-RAN Centralized Unit (O-CU), or in other remote nodes such as the near-real-time RAN intelligent controller (RIC), non-real-time RIC, or service management and orchestration, depending on the latency requirements of the use case.

Understanding the intelligent RAN architecture



The O-RAN Alliance has defined several use cases and an architecture for building open, intelligent RANs. The architecture includes the RIC, as shown in Figure 1. Alongside the O-RAN Radio Unit (O-RU), O-DU and O-CU, this architecture includes:

- **Near-real-time RIC (near-RT RIC):** This services microservices-based xApps, which can be used for use cases with a latency tolerance of 10 milliseconds to 1 second. The near-real-time RIC is typically hosted at the telco edge or regional cloud.

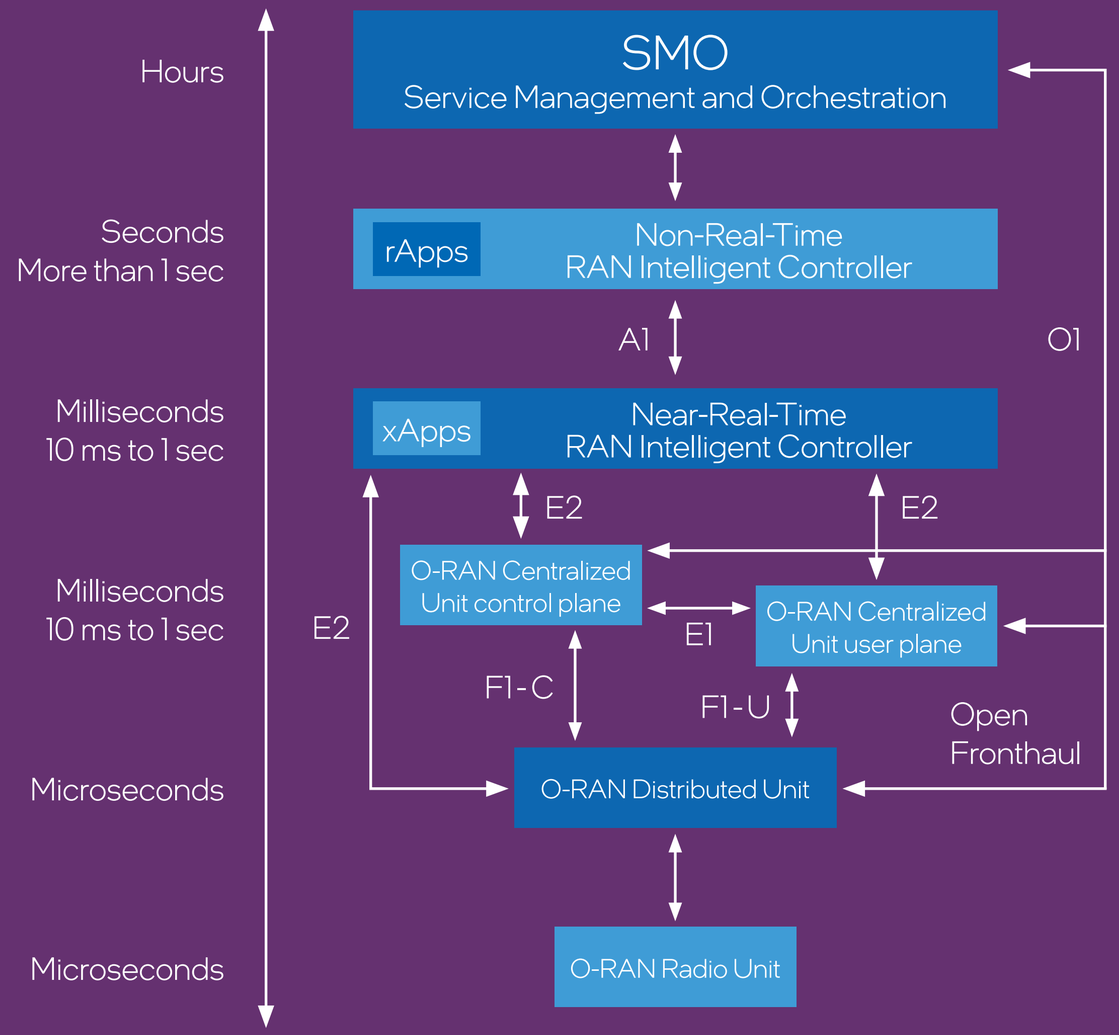


Figure 1: The O-RAN Intelligent RAN Architecture, showing the different functions, the latency for each one, and standardized interfaces between them for data collection and policy communication.

- **Non-real-time RIC (non-RT RIC):**
  This services microservices-based rApps, which are used for use cases with a latency tolerance greater than 1 second. The non-real-time RIC is hosted centrally in the network.

- **Service management and orchestration (SMO):**
  The SMO has long been part of the network, but it can now support the rApps and xApps by processing RAN data, for example, by training AI models.

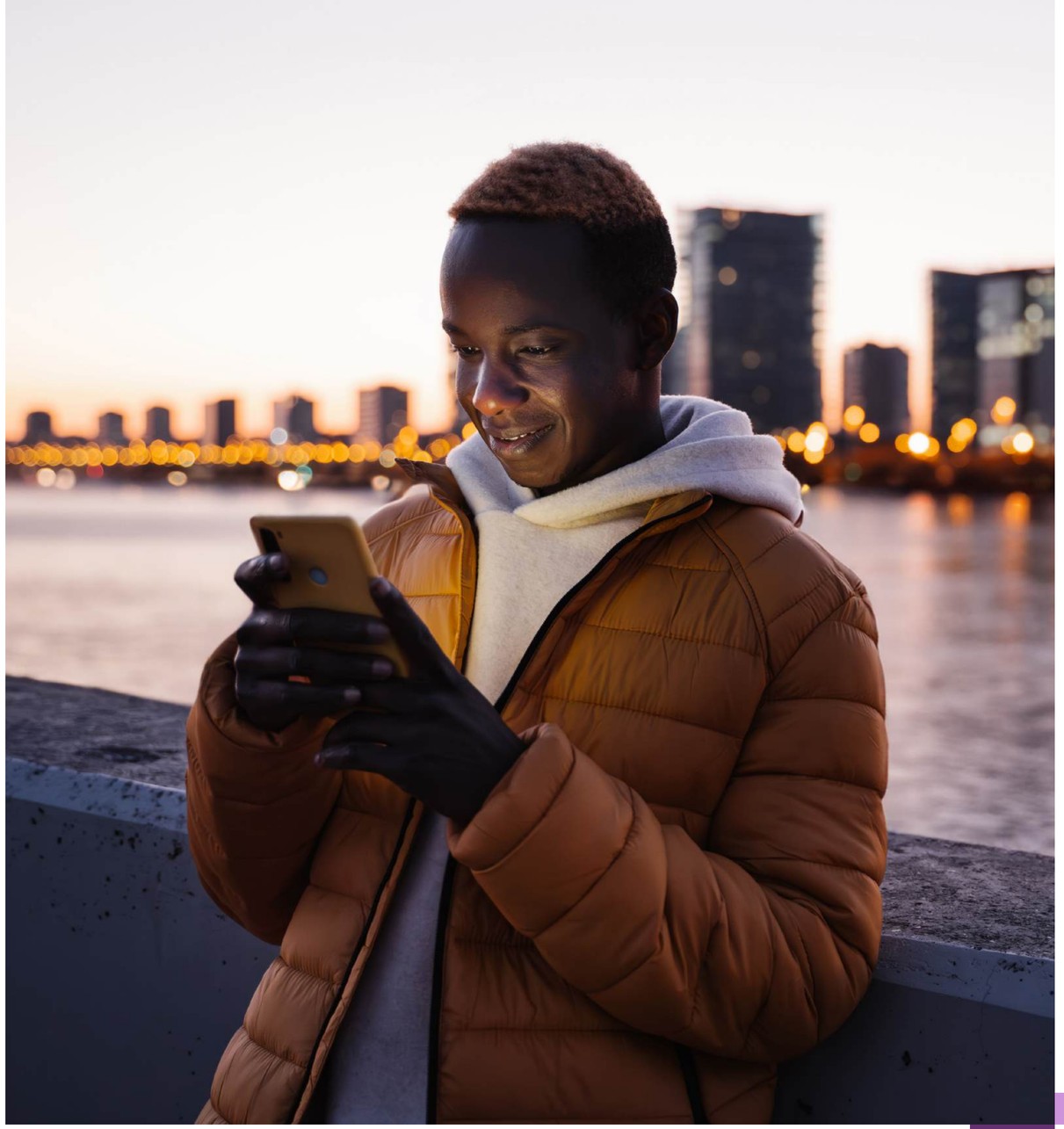
xApps and rApps are typically AI/ML-based models, although they need not be.

For use cases requiring sub-10 millisecond latencies, AI/ML models must be deployed closer to the edge, such as in the O-DU. In a fully virtualized RAN, AI can augment the functions of Layer 1 or Layer 2 of the RAN stack executing in the O-DU.

The O-RAN Alliance specification defines standard interfaces to transfer data and communicate policies (also shown in Figure 1). Real-time data includes such things as signal quality, telemetry data on the network's performance, and user throughput requirements. The scope of data required will continue to evolve as AI models are developed for both current use cases and future use cases.

A fully virtualized RAN can most flexibly adapt to evolving data needs from new use cases. Virtualization allows AI models to be added to augment or replace existing functions in the RAN stack pipeline for better performance and efficiency, and enables new business models. The simple programming model for general-purpose processors allows rapid and efficient network evolution, including with AI. AI algorithms can run on the same hardware as the RAN functions without an accelerator.

In many use cases, applications can use AI to process historical or current data. The resulting decisions are returned to the RAN functions nearer the radio as policies, traveling through the O-RAN standard E2 interface.



Let’s consider MIMO beam management from the non-real-time RIC. Beam management is a way to target the signal at a cluster of users. A simple example of this is an office with a restaurant below. During the day, the office on the upper floors is busy, while the restaurant on the ground floor is quiet, and at night, the restaurant is full. Beam management would allocate the beam with the radio frequency (RF) signal quality maximized upstairs during the day and downstairs after work. Typically, the network could adjust the beam on a schedule according to typical user patterns, but the non-real-time RIC would use information from the RAN to optimize beam management using current demand.

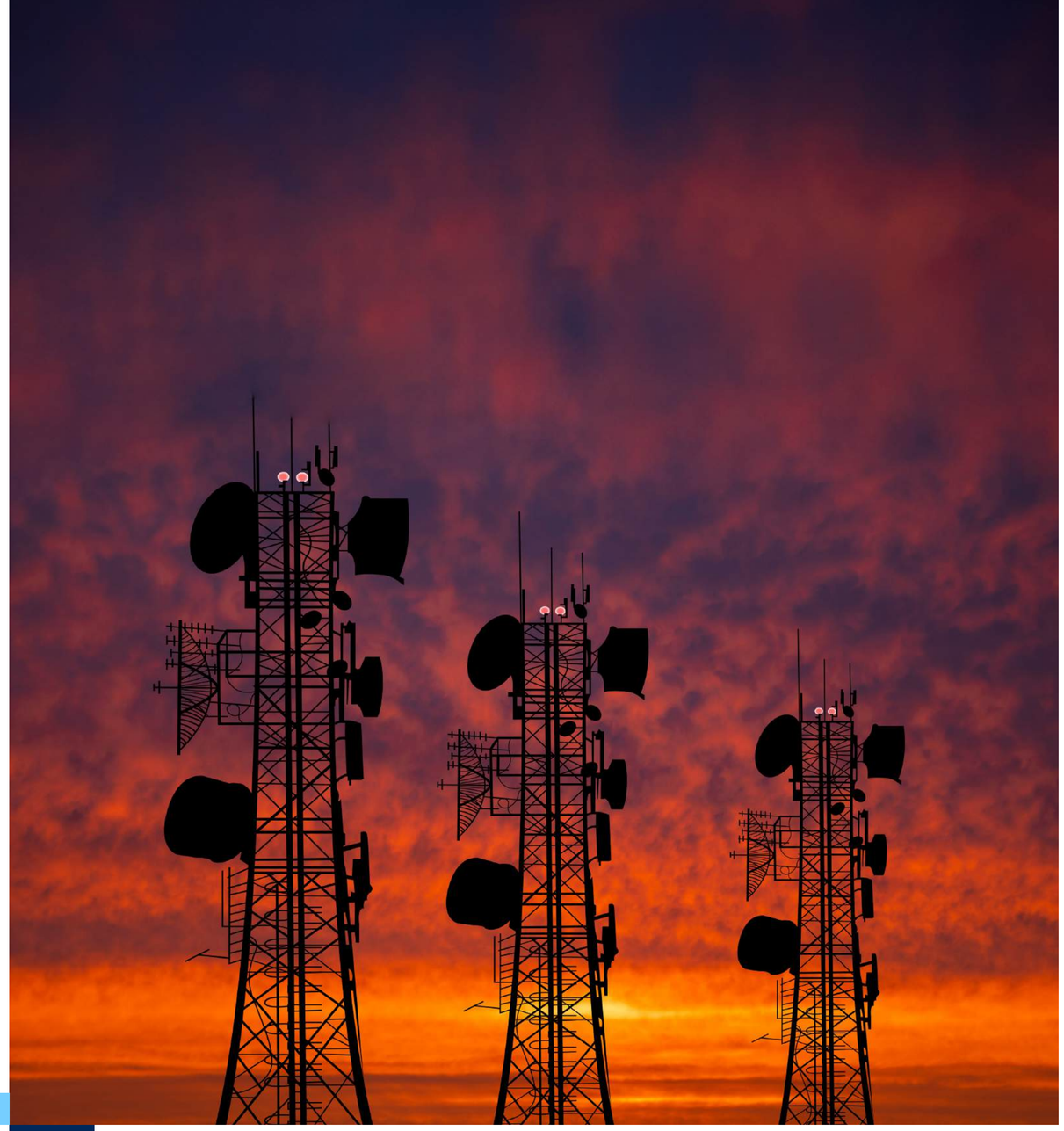
Table 1 shows some of the use cases for AI in the RAN and the typical location of the related models.

Location	AI RAN application
Non-Real-Time RIC	▪ Energy saving (switch cells off/on) ▪ Radio optimization to improve cell coverage and capacity
O-CU or Near-Real-Time RIC	▪ Anomaly detection for RAN key performance indicators (KPIs) ▪ Slice service level agreement (SLA) assurance using an xApp to predict performance indicators and avoid latency violation ▪ Power saving with dynamic CPU frequency (P-state) ▪ Multi-access traffic steering to improve throughput ▪ Connection management for handover to improve user throughput, load balancing, and coverage
O-DU	▪ Scheduler parameter optimization for better Quality of Experience (QoE) ▪ Link adaptation to improve cell throughput ▪ User selection to improve spectral efficiency ▪ Power saving (using C-states to put cores into microsleep states) ▪ Channel estimation to improve cell throughput

Table 1: AI use cases in the RAN

Demonstrations of AI in the RAN

We’ll now explore some demonstrations of AI in the RAN that Intel has enabled, working with other companies in the telco ecosystem.



Saving energy through dynamic power management

The RAN is responsible for 73% of the energy in the mobile network[3], and the O-RU accounts for most of that. Capgemini and Intel worked together to create an AI-enhanced rApp to reduce the power used by the O-RU.

The rApp monitors the energy consumption of the RAN nodes. It then uses a time-series machine-learning algorithm to forecast future energy consumption, carbon emissions, and load on the RAN. Input data includes wireless resource usage, service loads, the number of mobile users, and weather information. The load prediction uses data from a RAN node and its neighboring RAN nodes.

The solution was built and optimized using the Chronos software for scalable time-series analysis, part of the BigDL framework developed by Intel. The hardware is based on 3rd Generation Intel® Xeon® Scalable processors.

The rApp uses the predicted future load to apply energy-saving measures. These include:

- Switching off radio sites when they are not needed.
- Switching off high band carriers in a multiband cell when not required. Users are moved to low-band carriers (see Figure 2).

Capgemini tested the solution across twenty 4G and 5G radio sites with 150 cells. It could switch off up to 120 cells during the low-traffic period.

The algorithm uses closed-loop automation to improve the accuracy of the energy-saving decisions over time.

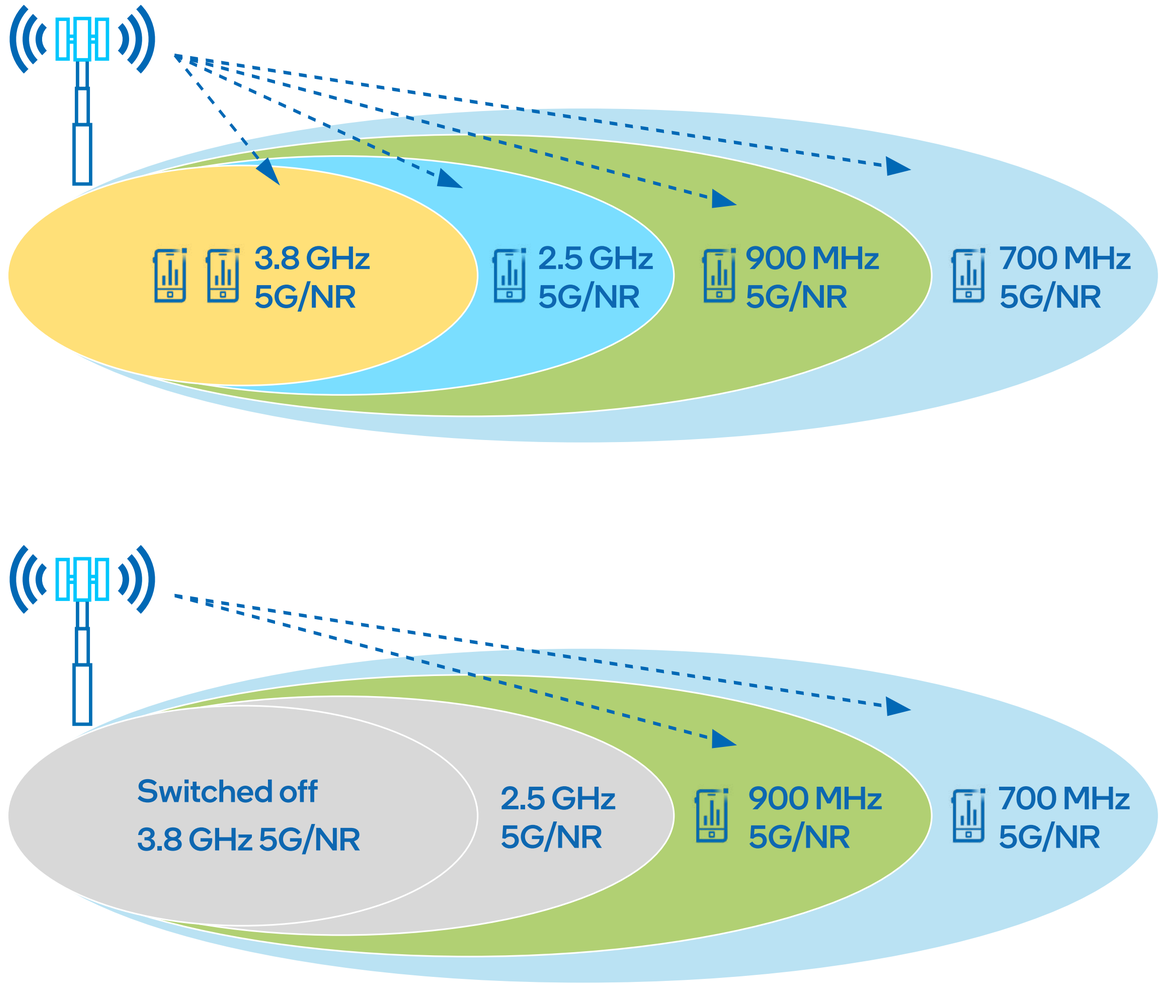
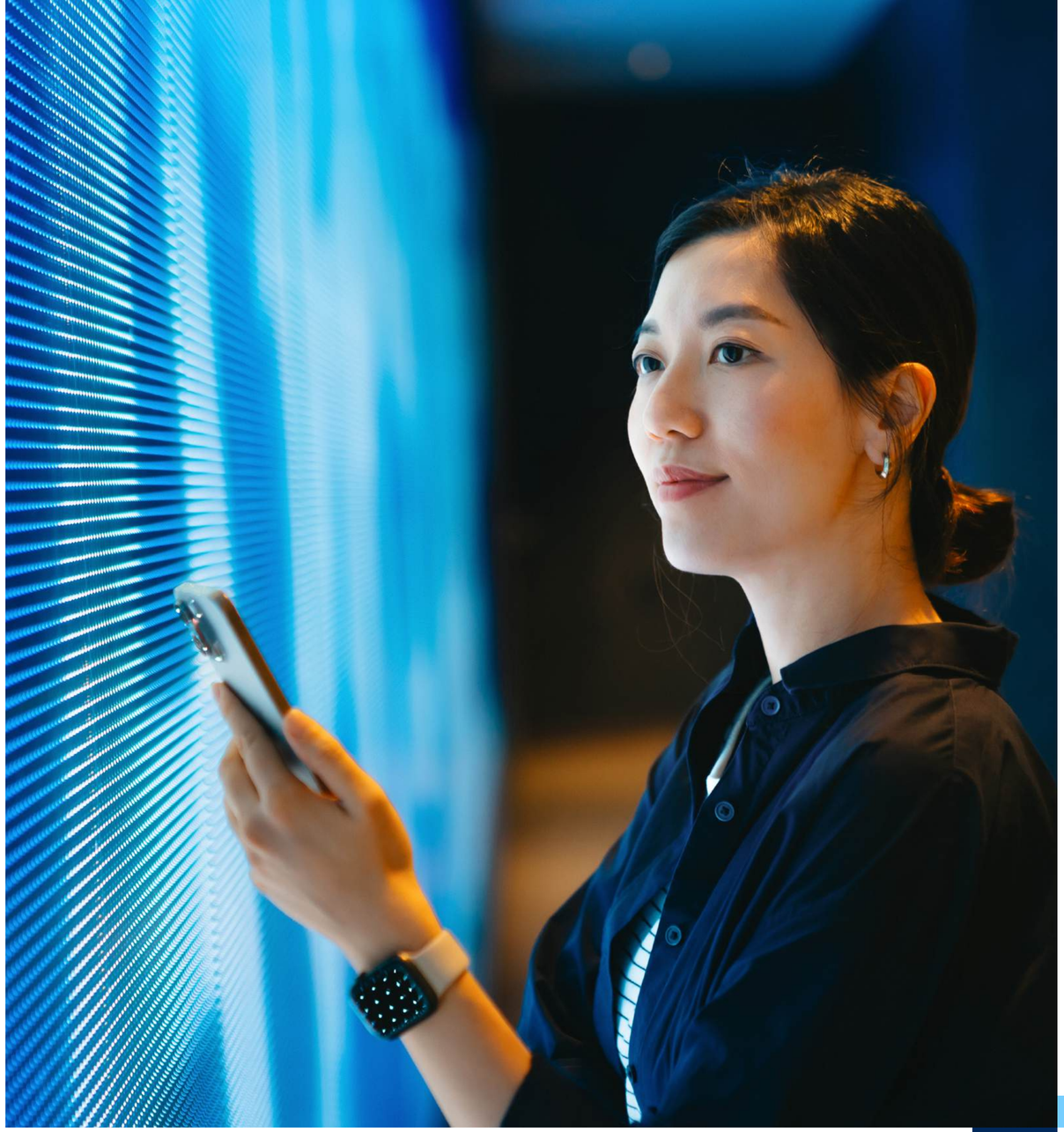


Figure 2: In times of high traffic (top image), all carriers are on. High-band carriers are switched off when traffic is low, and users are moved to low-band carriers.

Improving spectrum efficiency

In another project, Capgemini deployed its NetAnticipate5G and RATIO O-RAN platform to improve spectrum efficiency through advanced AI and machine learning (ML) techniques. This platform forecasts and assigns the appropriate MCS (modulation and coding scheme) values for signal transmission by predicting user signal quality and mobility patterns.

In this way, the RAN can intelligently schedule MAC resources to achieve up to 40% more accurate MCS prediction and yield up to 15% better spectrum efficiency.[4] As a result, it delivers faster data speeds, better and more consistent QoE to subscribers, and robust coverage for use cases that rely on low latency connectivity, such as robotics-based manufacturing and vehicle-to-everything (V2X) connectivity.



“We gathered and utilized over one terabyte of data and conducted countless test runs with NetAnticipate5G to fine-tune the predictive analytics to meet diverse operator requirements,” said Walid Negm, Chief Research and Innovation Officer at Capgemini Engineering. “In short, machine learning can be deployed for intelligent decision-making on the RAN without any additional hardware requirement. This makes it cost-efficient in the short run and future-proof in the long run as we move into cloud-native RAN implementations.”

Download the white paper: Intelligent 5G L2 MAC Scheduler.

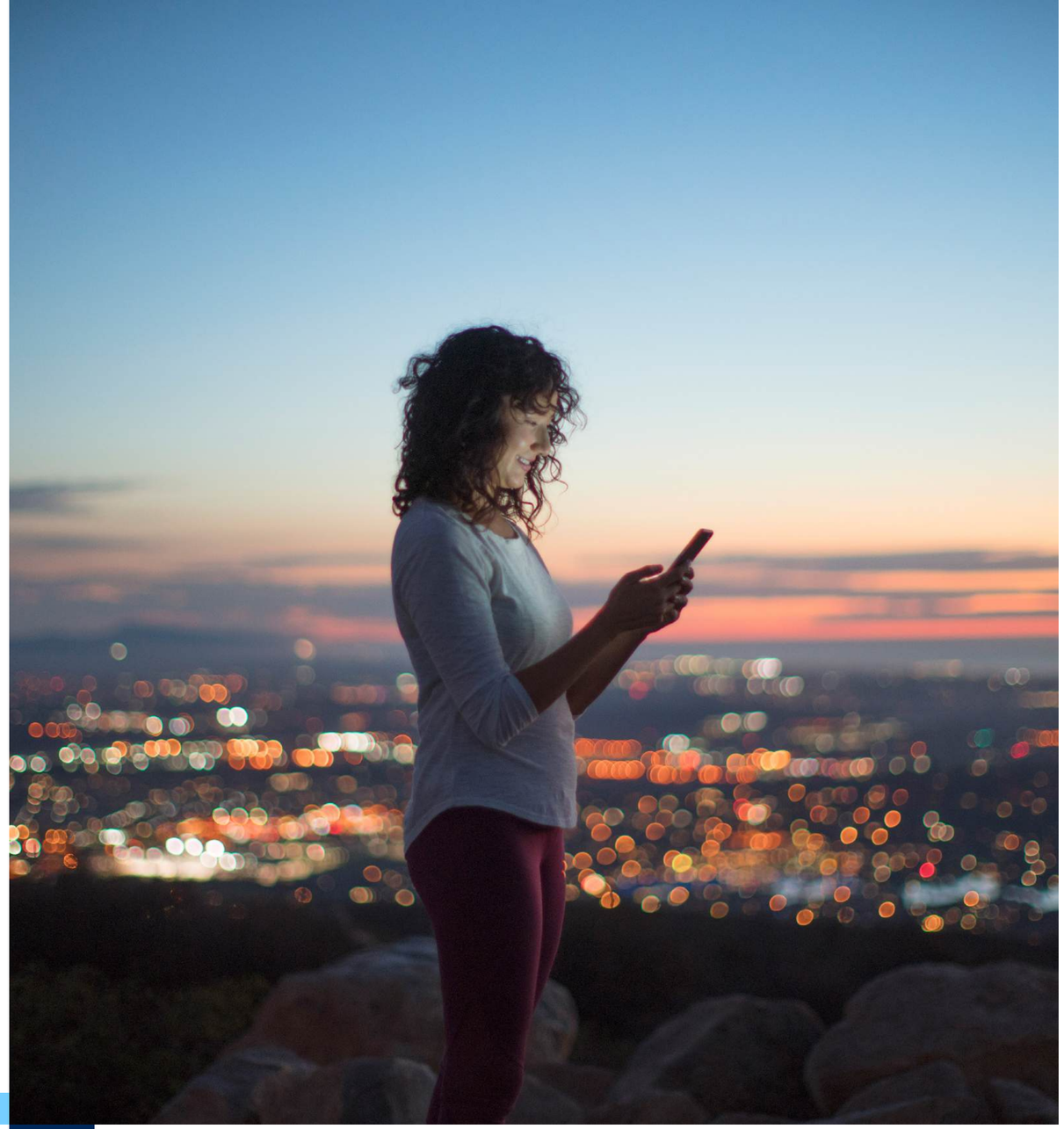
Improving MU-MIMO efficiency

Aira Technologies has demonstrated an xApp to improve channel estimation and channel prediction performance, helping to optimize downlink throughput and range. Accurate channel prediction helps minimize inter-user interference, which improves multi-user multiple-input multiple-output (MU-MIMO) performance.

The xApp uses machine learning and is assisted by the Intel® FlexRAN™ reference software, which fully implements O-DU Layer 1 and above. Intel FlexRAN software enables the xApp to access the physical Layer 1. The xApp is designed to work in concert with the VMware RIC platform and uses Capgemini's L2 and L3 software platforms. The hardware Aira Technologies is using is based on Intel® Xeon® processors.

“Radio Access Networks (RAN) operate in varied network conditions and radio environments,” said Ravikiran Gopalan, Founder and CTO of Aira Technologies. “ML presents a natural framework to classify these conditions accurately and process the RAN signals optimally for each of these conditions. We are seeing tremendous MU-MIMO throughput gains from our ML-based channel prediction xApp, and we are working on applying our ML framework to other RAN functionalities.”

Read the Aira press release: xApp improves 5G MU-MIMO efficiency and throughput using AI.



Improving resource utilization of O-DU upper PHY software

DeepSig has developed embedded software that replaces multiple 5G NR signal processing algorithms with a precisely designed deep neural network (DNN).

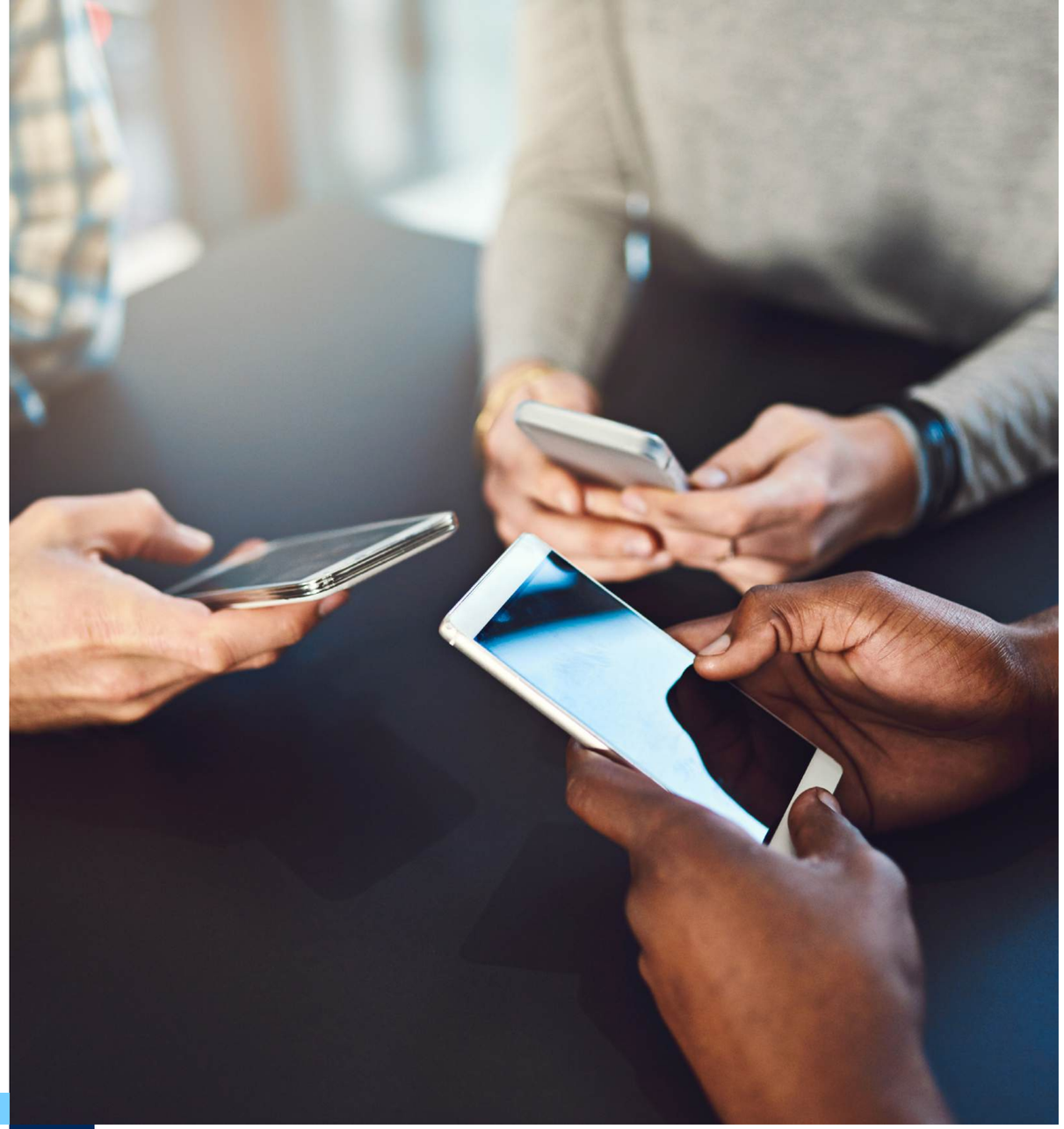
This approach potentially requires less computation while significantly improving network capacity and resilience to interference by learning the real-world characteristics of the local wireless environment where the O-RU operates. These improvements reduce both capital expenditure and operating expenses. DeepSig and Intel collaborated to bring this transformational AI software to market as part of the Intel FlexRAN software suite.

DeepSig's 5G AI embedded software provides drop-in replacements to the physical uplink shared channel (PUSCH) channel estimation routines (for standard MIMO), and to SRS channel estimation and pre-coding routines (for Massive MIMO). DeepSig's 5G AI software components can be leveraged by existing Intel FlexRAN software for O-DU vendors without additional hardware or software stack changes.

By using machine learning for Layer 1 processing, less compute is required to process the uplink (PUSCH) signals in standard MIMO configurations, increasing the number of sectors per server and reducing their cost of operation.

The ML-driven Layer 1 processing improves the signal-to-interference-plus-noise ratio (SINR). This enables increased throughput and coverage, optimizing the value and utilization of costly spectrum licenses and band allocations. SINR improvements can also enable bandwidth increases, reduced user traffic latency, and a smaller interference margin for cell planning.

Download the white paper: Amplifying 5G vRAN performance with artificial intelligence and deep learning.



Improving traffic steering

Mobility robustness optimization (MRO) dynamically manages the network performance of handovers to improve end-user experience and increase network capacity. The aim is to eliminate radio link failures and reduce unnecessary handovers. MRO works by automatically adjusting handover boundaries based on performance indicators. MRO minimizes human intervention in the network management and optimization tasks, thus reducing operator expenditures.

Handover decisions made by the MRO capability only consider cell-level statistics in fine-tuning the handover thresholds. Traffic management solutions usually relocate users among cells, relying heavily on user equipment (UE) feedback in the measurement report. The statistical characteristics of the radio network and information about the UE behavior are not currently used to enhance the network and user experience performance.

Mavenir has demonstrated the ability to proactively manage specific user traffic across different cells and access technologies, using information about the UE behavior and the radio network. A mobility management rApp provides policy recommendations for handovers and sends them to the near-RT RIC for policy enforcement.

The traffic steering xApp hosts a reinforcement learning (RL) agent. From the O-CU and O-DU, the xApp receives context and state information about the UE and the serving cells, the UE's layer 3 radio resource control (RRC) information for neighbor cells, and neighbor cell context information. The RL agent in the xApp then generates a UE-specific handover control action that optimizes the decision of the target cell for the UE. This decision maximizes a given key performance indicator for the UE, such as throughput or latency. In this way, traffic steering enables operators to provide an intent-driven UE-specific service.

The outcomes include:

- Better resource utilization of RIC compared to traditional self-organizing network (SON) and radio resource management (RRM) algorithms.
- Average 50% improvement in throughput with RIC compared to SON and RRM.[5]
- Average 52% improvement in UE spectral efficiency with RIC compared to SON and RRM.[5]

The reinforcement learning models to make these RAN performance gains are accelerated by Intel® Optimization for TensorFlow, which takes advantage of Intel® AI Engines for the 4th Generation Intel® Xeon® Scalable processor. The reinforcement learning AI performance results below meet Mavenir's AI performance requirements:

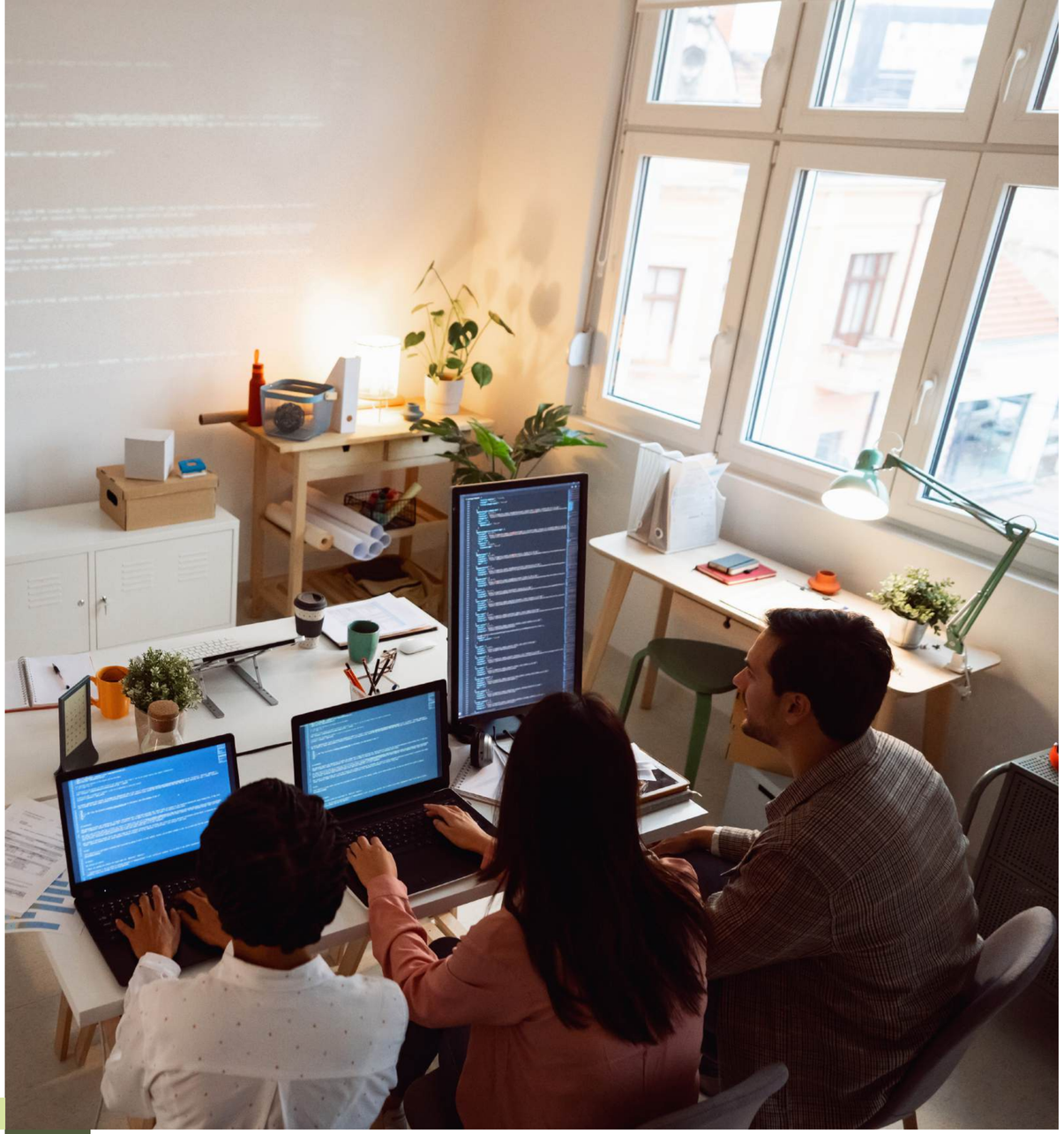
- **AI training latency:** 7.72ms with single CPU core[5]
- **AI Inference latency:** 0.65ms with single CPU core[5]

Download the white paper: Open vRAN radio intelligent controller expands RAN Capability.

Accelerating AI on general-purpose processors

Hardware performance is important for AI-based applications. The telco network has strict latency requirements for some functions, and associated AI applications should be processed close to the wireless transmission. In a virtualized DU, the same general-purpose processor that supports Layer 1 and 2 functions can also run the AI application.

4th Gen Intel Xeon Scalable processors include several features that accelerate AI models. These include Intel® Advanced Vector Extensions (Intel® AVX) for vRAN, support for lower-precision calculations for faster AI training and inference, and Intel® Advanced Matrix Extensions (Intel® AMX).



To help developers take advantage of processor features such as these, Intel has a strong ecosystem of software support. Intel has released optimized AI libraries, such as Intel® oneAPI Deep Neural Network Library, Intel® oneAPI Math Kernel Library, and Intel® oneAPI Data Analytics Library.

The Intel® oneAPI AI Analytics Toolkit provides tools and frameworks to accelerate the end-to-end analytics and AI pipeline on Intel® architecture processors. It provides an Intel-optimized framework for TensorFlow and PyTorch with low-precision tools for high-performance training and inference. It accelerates data preprocessing and machine learning workflows with the Python packages Modin, scikit-learn, and XGBoost.

Other AI acceleration features include support for Chronos, which enables developers to build time-series forecasting AI models, including data processing and feature engineering. Intel oneAPI AI Analytics Toolkit includes built-in AI models, a user-friendly Application Programming Interface (API), and AutoML for automatic hyperparameter optimization in a distributed architecture. AutoML enables developers to train and tune highly accurate AI models.

Conclusion

While operators have already started deploying AI in select use cases, there is huge potential, and a number of demonstrations have already shown what is possible.

The benefits of introducing AI in several layers in the RAN have been demonstrated. The flexibility of a fully virtualized RAN makes it possible to leverage AI capabilities by enhancing parts of the RAN processing pipeline.

As operators introduce virtualized RAN to their networks today, they are laying the foundations for future AI-based enhancements.

Intel® Xeon® processors include built-in acceleration for AI training and inferencing and are supported with tools, frameworks, and software for rapid implementation. Ecosystem partners, including Capgemini, Aira, DeepSig, and Mavenir, use them to enable their AI-based RAN technologies.

Learn more

- Amplifying 5G vRAN performance with artificial intelligence and deep learning
- White paper: Intelligent 5G L2 MAC Scheduler
- Accelerate AI workloads with Intel® AMX

1. European mobile data traffic will triple in next five years, GSMA, November 2023
2. 2023 annual survey highlights, CTIA, November 2023
3. GMSA Intelligence, Going green: Benchmarking the energy efficiency of mobile, June 2021
4. Industry's first machine learning-based ran application boosts spectral efficiency by 15%, Capgemini, 30 June 2021
5. System: Intel Corporation ArcherCity; Number of nodes: 1 node; Baseboard: Intel Corporation ArcherCity; Chassis: Rack Mount Chassis; CPU Model: Intel® Xeon® Platinum 8480+; Microarchitecture: 4th Gen Intel® Xeon® Scalable Processor; Sockets: 2; Cores per Socket: 56; Hyperthreading: Enabled; CPUs: 224; Intel Turbo Boost: Enabled; Base Frequency: 2.0GHz; All-core Maximum Frequency: 3.0GHz; Maximum Frequency: 3.8GHz; NUMA Nodes: 2; Prefetchers: L2 HW, L2 Adj., DCU HW, DCU IP; PPINs: 08b7d81fa52198c3,08b0d21f7f5ec0df; Accelerators: DLB:2, DSA:2, IAX:2, QAT (on CPU):2, QAT (on chipset):0; Installed Memory: 256GB (8x32GB DDR5 4800 MT/s [4800 MT/s]); Hugepagesize: 2048 kB; Transparent Huge Pages: madvise; Automatic NUMA Balancing: Enabled; NIC: 1x Ethernet Controller I225-LM; Disk: 1x 3.6T Samsung SSD 870 QVO 4TB; BIOS: EGSDCRB1.86B.0090.D03.2210040151; Microcode: 0xab000310; OS: Ubuntu 22.04.1 LTS; Kernel: 5.15.0-43-generic; TDP: 350 watts; Power & Perf Policy: Performance; Frequency Governor: performance; Frequency Driver: intel_pstate; Max C-State: 9; Tensorflow: 2.11; Tested AI Workloads: Mavenir reference learning training and inference for traffic steering xAPP. Tested by Intel as of 04/14/2023

1223/JH/CAT/PDF 358291-001EN