The future of the industrial AI edge is cellular

Xenofon Foukas
Microsoft
Cambridge, UK
xefouk@microsoft.com

Bozidar Radunovic
Microsoft
Cambridge, UK
bozidar@microsoft.com

Abstract

Ensuring reliable and high-bandwidth wireless connectivity and local processing at the edge are crucial enablers for emerging industrial AI applications. In this work, we argue that the recent trends in cellular networking make the technology the ideal connectivity solution for these applications, due to its virtualization and support for open APIs. We foresee the emergence of a converged industrial AI edge encompassing both compute and connectivity, in which application developers leverage the API to implement advanced functionalities. We demonstrate the usefulness of this approach through a case study evaluated on an enterprise-grade 5G testbed deployed in our lab.

CCS Concepts

• **Networks** → **Network architectures**; **Mobile networks**; • **Applied computing** → **Enterprise computing**.

ACM Reference Format:
Xenofon Foukas and Bozidar Radunovic. 2025. The future of the industrial AI edge is cellular. In *The 26th International Workshop on Mobile Computing Systems and Applications (HOTMOBILE '25), February 26–27, 2025, La Quinta, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3708468.3711887

1 Introduction

The industrial enterprise is undergoing a major digital transformation. The recent advances in AI/ML have led to an explosion of use cases (see Table 1) which are envisioned to greatly simplify common tasks, drive down the costs and create safer environments [25]. Many of these emerging applications are characterized by their need for processing at the edge, making the last hop connectivity between the devices and the edge infrastructure a key component (Section 2).

Considering the reliance of emerging industrial AI applications on high-bandwidth modalities for inference (e.g., video, audio, RF), there only exist a limited set of options that can accommodate their needs. Wired connectivity is most often too inflexible and expensive to deploy. At first glance, Wi-Fi appears to be a compelling candidate, given its high capacity, low cost and low complexity. But it fails to provide the reliability that industrial applications require. On the other hand, it has been long advocated that cellular connectivity can fill these gaps. However, so far, its high cost and complexity made it suitable only for niche use cases.

In this work, we want to challenge this perception. We argue that with the recent trends, the 5G and beyond cellular technologies have the potential to eliminate the high barrier-to-entry, offer complexity similar to Wi-Fi, superior connectivity and more features, at a comparable cost. We observe that the key components in this vision are the softwarization of the cellular stack and the standardization of APIs. Furthermore, we argue that this connectivity shift will see a convergence to a single edge AI compute and communication architecture. To back up our claims, we perform a case study by building a few of these use cases and evaluating them on an enterprise-grade 5G testbed [8] deployed in our lab in Cambridge, UK.

Vertical industry	AI-driven application example
Manufacturing	Intelligent factory conveyor belt, predictive maintenance, humanoid robot workers
Logistics	Detection and replenishment of missing stock with autonomous mobile robots
Airports/Stadiums/Malls/ Enterprise campuses	Video analytics for security or marketing purposes
Ports/Mines/Utilities	Automation of dangerous tasks (e.g., freight lifting and drilling)
Agriculture	Crop quality assurance checks using video feeds from robots or drones

Table 1: Examples of verticals and applications that are being transformed by AI.

2 Characteristics of the industrial AI edge

The recent AI advances are unlocking new opportunities for the enterprise. Recent studies show that more than 14 million industrial sites are being transformed or are likely to be transformed in the coming years, driven by the emergence of applications that heavily rely on the use of AI [25]. As shown in the examples of Table 1, such applications and industries can be very diverse. For example, some applications might be targeting large indoor spaces (e.g. warehouses and factory floors), others might be targeting large outdoor spaces (e.g. ports, enterprise campuses) or hard-to-reach locations (e.g. agricultural fields, mines, and utility infrastructure), while others might be focusing on dense indoor or outdoor spaces (e.g. stadiums, shopping malls). Regardless of the exact environment, all the aforementioned applications are characterized by the following requirements:

Reliable high-bandwidth connectivity – Many new AI applications rely on high bandwidth data sources (e.g., video, audio, lidar, RF sensing) and multi-modal inference, requiring high-bandwidth connectivity. These applications also need to meet QoS guarantees in terms of throughput and latency, as well as to experience uninterrupted coverage. The lack of reliable connectivity could result in the performance degradation of applications (e.g., dropped frames or low resolution for video analytics) or, more seriously, safety risks (e.g., a robot causing accident due to spotty connection or lost packets).

Processing at the edge – The aforementioned applications will require edge processing capabilities, mainly due to limited backhaul bandwidth or due to the lack of reliability of the network link towards the cloud [23]. For example, some critical applications in the context

HOTMOBILE '25, February 26–27, 2025, La Quinta, CA, USA

ACM ISBN 979-8-4007-1403-0/25/02
https://doi.org/10.1145/3708468.3711887

of manufacturing or ports might require uninterrupted operation, as any interruption could translate to significant financial losses. Others, like video analytics and agriculture, might require processing at the edge, due to the high volume of data they produce, which makes the communication to the cloud very costly. Further factors that favor edge deployments are privacy and latency.

3 Connectivity challenges

Connectivity plays a crucial role in enabling the future industrial AI edge. Given the diverse locations, the morphology of the terrain, and the device density served by enterprise edges, the deployment of wired connectivity solutions is prohibitive, making wireless connectivity the only viable option. There are several wireless technologies used in the industry today [6]. Many are tailored for low-bandwidth use, while Wi-Fi has emerged as the de-facto wireless solution for high-bandwidth connectivity. This widely accepted perspective is driven by the characteristics of Wi-Fi in terms of its low cost, high bandwidth and low deployment complexity. We argue that while this choice is preferable for more traditional enterprise use cases (e.g., people connectivity in small offices), it presents several critical limitations with respect to the emerging industrial AI use cases:

Reliability – Many industrial applications require stable throughput and latency [15]. We argue that, while WiFi can provide similar results to 5G in the average case in terms of latency and loss, it can have a long tail, due to its operation in unlicensed spectrum and the presence of interference from other co-existing networks. To make things worse, the use of Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) means that dense environments with high contention amongst devices can lead to further degradation of performance. Several recent studies have highlighted these issues experimentally. For example, the study in [22] has shown that WiFi can have uplink latencies of up to 100ms, when co-existing with just a single WiFi 5 network on the same channel. The work in [31] observed that such latencies can be even higher and can reach more than 200ms. This, in addition to the low/variable throughput of WiFi (c.f. [22, 31] and also Section 6), lead to KPIs that are not acceptable for mission-critical workloads (e.g., robotic warehouses [14]). On the other hand, 5G can provide predictable and stable latency, due to its operation under licensed spectrum, and its centralized scheduling (e.g., as shown experimentally in [7] with the use of network slicing).

Coverage – Today, Wi-Fi is more often deployed in the higher 5 GHz and 6 GHz bands, in contrast to the 2.4 GHz bands, because it offers superior capacity. The downside is that its signal propagation is significantly shorter. This poses a critical challenge in many industrial settings, which span large areas, and where feasible access point deployment locations are often scarce [11]. While one could consider the 2.4 GHz spectrum for longer range, its channels are much more utilized [10], leading to reliability issues. Even in cases where the densification of the network is possible, the addition of more Wi-Fi access points introduces more coordination complexity, the mitigation of which is still an open research problem [28].

These limitations of Wi-Fi have been widely recognized by both the industry and academia as key pain points. As a more promising alternative, it has been long argued that cellular connectivity has the potential to replace Wi-Fi in industrial settings, due to its increased coverage, improved reliability and support of several Radio Access Technologies tailored for different use cases [11, 25]. However, despite this long-standing promise, cellular technologies have so far not met any real success in the enterprise setting due to their high deployment complexity and cost, and the requirement for specialized telecom expertise to manage them, which makes them suitable only for a very niche and limited set of use cases.

4 Cellular in the spotlight: A paradigm shift for the industrial AI edge

We argue that the recent advances in the context of cellular networks, combined with the emerging needs for AI at the edge, open up a unique opportunity for cellular (i.e., private 5G) to become the standard technology of the industrial edge of the future. Here, we discuss our observations that drive this claim:

Simplification through software – 5G and beyond mobile networks have undergone a paradigm shift over the last decade, converting the cellular infrastructure from purpose-built specialized embedded devices to virtualized network functions. Even the most involved part of the infrastructure, the radio access network (RAN), is today implemented fully in software, on top of commodity hardware, with software-based stacks already in the mainstream [7]. While this transformation has been mainly driven by the demand of the operators in the telco space to reduce their costs, it has also created the opportunity for enterprise solution vendors to build cheaper and simplified versions of the cellular stacks, tailored for enterprises and the industry. Such software-based solutions have been designed to hide most of the cellular complexity (e.g., thousands of 3GPP parameters) that has always daunted the non-experts. A case in point are several high-quality open source projects, such as srsRAN and Open5GS, which provide reliable and efficient [2] containerized software implementations of cellular components, simplified for these use cases. As we show in our case study (Section 6), we believe that their performance and stability finally make them (and similar commercial enterprise solutions) ready for the prime time.

The softwarization of cellular solutions also means that they can now be treated as yet-another-set-of-apps that can be managed and deployed on a virtualized infrastructure (e.g., in a kubernetes deployment). This in turn means that we can easily roll out new features on-the-fly, significantly increasing the lifespan of the HW infrastructure. For example, one could move from 5G to 5G advanced with just a simple software upgrade (in contrast to WiFi 6 vs WiFi 7, where all the access points would have to be replaced).

Shared commodity hardware – The transformation of the cellular network functions to software also means that both cellular software and edge AI apps can now be collocated on a single platform and over the same hardware, which greatly simplifies their manageability. Commodity servers with general purpose processors (e.g., x86 or ARM) and GPUs can already be used to run both cellular software and AI edge workloads on the same processor die (e.g. Intel Xeon Sapphire Rapid EE and OpenVINO [20, 21] and Nvidia GH200 with the Aerial SDK [24]). This means that there is no longer a need to invest separately for the compute of the cellular infrastructure,

significantly driving down costs. This cost reduction can be further enhanced through the statistical multiplexing of the cellular and edge AI software at runtime, driven by the observation that the cellular infrastructure is commonly underutilized (<50% utilization) even at peak hours [4, 19]. We demonstrate the benefits of this approach in Section 6, through a proof-of-concept of RAN compute sharing that we have implemented in our lab, which allows us to run both the RAN and ML workloads over the same CPU cores.

Flexible and open interfaces – Industrial AI applications are data-driven and consist of numerous niche use cases, each with its own set of requirements, in terms of coverage, QoS guarantees, etc. For example, a video analytics edge application may leverage network APIs to collect real-time RAN data and enforce deadline-sensitive packet scheduling to achieve low latency (cf. [30]). It is thus key to expose flexible interfaces to allow application developers to tap into the connectivity fabric, to collect data and to tailor it to the use case in mind. This can significantly drive down the cost by leading to the simplification of the network stack, allowing customers to only pick the features that they care about, and by opening the ecosystem to new players, increasing innovation and competition.

We argue that cellular is better positioned to expose diverse APIs compared to Wi-Fi. Most low level Wi-Fi functionalities, like signal processing are implemented in hardware and are much harder to modify. This has led to a few proprietary APIs, only provided by major Wi-Fi vendors, offering restricted functionality and interoperability. This is in stark contrast with the open and inter-operable interfaces of cellular, which offer a lot more flexibility for innovation. For example, O-RAN APIs are standardized, have received careful scrutiny with regards to their security and privacy (through a dedicated security working group), they are widely accepted in the cellular industry and can be easily extended due to virtualization. As an example, we show in Section 6 that, by tapping into the low-level open fronthaul interface of the RAN that connects the RUs to the base stations, we are able to implement a prototype distributed MIMO middlebox solution that significantly extends the network coverage, without the need to deploy more cells and deal with complex problems like mobility and handovers. We also discuss other examples, like the use of network slicing for providing service performance guarantees.

Commoditized radio access front-end – Another key transformation in the cellular space is the recent availability of shared spectrum in many countries [16]. This, in conjunction with the standardized front-haul interface, led to the increasing availability of 5G radio units from smaller vendors (e.g., Foxconn, VVDN, Benetel) and is significantly closing the gap in the cost between deploying Wi-Fi and cellular networks (see our cost analysis in Section 6).

It should be noted that the above benefits of private 5G are already being acknowledged by major industrial players. For example, Tesla is rolling out private 5G at scale in their gigafactories in Berlin and Shanghai [3]. Similarly, Airbus has announced the replacement of WiFi with private 5G in all their plants within the next 5 years [1].

5 The vision of converged compute and connectivity, with cellular and open APIs

Driven by the observations of Section 4, we now present our vision of a converged architecture for the future industrial AI edge. As illustrated in Figure 1, this architecture is based on commodity hardware

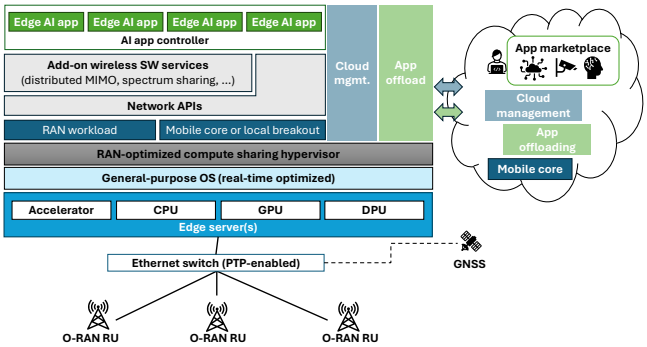
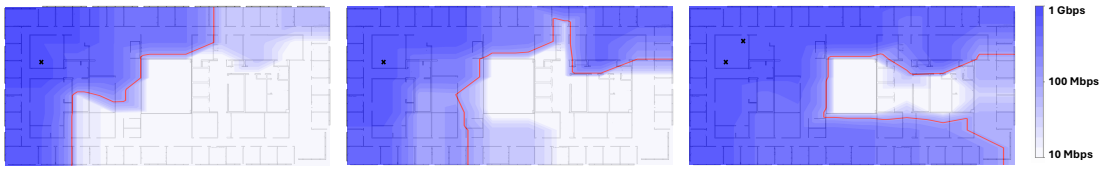


Figure 1: High level overview of the future converged industrial AI edge.

(radio units, switches and servers). The servers are equipped with processors and accelerators that can be leveraged by both cellular and AI workloads (e.g., CPUs with AI and vRAN instruction set architecture extensions and accelerators, GPUs, DPUs, etc.). They are operated using a platform software that includes a general purpose operating system (e.g., Linux), optimized for high-performance and real-time workloads. It also includes a remote management software and an AI controller orchestrating and executing the edge AI applications (e.g., along the lines of [5]). Given that the same platform will run both cellular software (RAN and all or part of the mobile core) and edge AI applications, the platform also includes a RAN-optimized compute sharing hypervisor (e.g. building on KVM or vGPU technologies) to enable the efficient and reliable statistical multiplexing of edge and cellular workloads. Finally, given that the multiplexing of workloads could occasionally lead to the saturation of the server compute resources, the platform also provides services for dynamically offloading edge AI applications to the cloud.

On the programmability front, the cellular software stack provides standard open interfaces that allow application developers to both tap into real-time network data, as well as improve and customize the connectivity layer for their use cases (e.g., via network slicing [17]). We envision that the cellular software will have minimum extra features in order to drive the cost down. It will be based on stripped-down versions of tier-1 macro products, or hardened versions of existing open source products or reference designs. Further customization will be done separately for each market, by leveraging third-party solutions built around the open interfaces (see the distributed MIMO example in Section 6). Finally, we expect to see a marketplace that will offer a plethora of off-the-shelf cellular software, network control applications and general-purpose edge applications (video analytics, LLM, etc), that can tend to the needs of different verticals.

It should be noted that our vision is well-aligned with several ongoing industrial efforts that have recently emerged in this space. For example, SoftBank and Nvidia recently announced AITRAS, which is a converged solution for running AI workloads at the 5G edge [26]. Similarly, Verizon recently announced a Mobile Edge Compute private 5G solution in collaboration with Nvidia, for deploying industrial real-time AI workloads [27]. While such efforts still lack advanced features like the ones outlined above (e.g., RAN-optimized hypervisors, advanced connectivity applications), we believe that they could be pivotal for triggering a paradigm shift for industrial connectivity.



((a)) WiFi ((b)) Regular 5G ((c)) 5G with distributed MIMO

Figure 2: Approximate coverage for Wi-Fi and 5G measured in our enterprise setting. Wi-Fi downlink and uplink are approximately the same. The red line shows the 100 Mbps limit. The white area offers less than 10 Mbps. The black X marks are radio deployments.

	5G		Wi-Fi	
Carrier:	3.5 GHz		5 GHz	
BW:	100 MHz		80 MHz	
	DL (4x4 MIMO)	**UL (SISO)**	**DL (2x2 MIMO)**	**UL (2x2 MIMO)**
DL	700 Mb/s		700 Mb/s	
UL		150 Mb/s		600 Mb/s
DL+UL	700 Mb/s	150 Mb/s	350 Mb/s	350 Mb/s
4 × DL	700 Mb/s		300 Mb/s	
4 × UL		150 Mb/s		300 Mb/s

Table 2: Throughput comparison of 5G and Wi-Fi with typical parameters with one and four mobile devices.

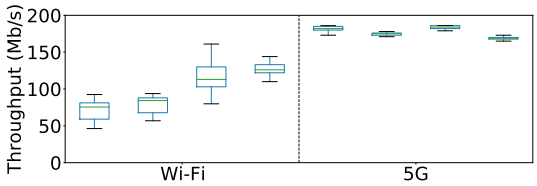


Figure 3: Throughputs for 4 phones over time continuously transmitting downlink, for the two technologies.

6 Case study and evaluation

To illustrate the points raised in Sections 4 and 5, we perform a case study in our 50.9m × 20.9m office building in Cambridge, UK. While we acknowledge that this is not the most typical example of an industrial AI deployment, we believe it is realistic enough in the context of this study. For the evaluation, we leverage the enterprise-scale 5G testbed presented in [8], comprised of HPE servers with Intel Xeon 6338N CPUs (32 cores) and Foxconn RPQN 7800 radio units. In contrast to the setup in [8], we replace the cellular software stack with open source containerized components, i.e., 5G base stations based on the srsRAN stack and a 5G core based on Open5GS. Both projects provide excellent documentation that allowed us to get everything up and running in just a few days, without prior experience. In our experiments we use four smartphones from two vendors (2x OnePlus N10 5G, 1x Samsung Galaxy A52s, 1x Samsung Galaxy S23).

Comparison of 5G against Wi-Fi – We begin by exploring the differences of private 5G over Wi-Fi in terms of performance, reliability and coverage. For this comparison, we use a Netgear Business WAX214 Wi-Fi 6 access point, typical for enterprise deployments. For fairness, we place the Wi-Fi access point right next to one of our Foxconn RUs at one side of our floor (black mark in Figure 2(a)).

Using this setup, we perform a UDP iperf experiment, to measure the throughput that we can achieve with each solution on the uplink and downlink, at close distance (~5 meters). The results can be seen at Table 2. As we can observe, the overall results are comparable in terms of performance, with some differences on the max throughput due to the differences in the exact configurations of each solution (e.g., use of MIMO or SISO for the uplink, etc). This similarity is expected, considering that both technologies are OFDMA-based.

The main difference is in scheduling. We observe that for a single uplink user, 5G offers lower capacity than Wi-Fi. This is because 5G uses fixed TDMA scheduling whereas Wi-Fi's CSMA can automatically adjust to uplink traffic only. However, CSMA is inefficient for multiple users [13]. We see that its aggregate throughput significantly drops (>50%) when four devices are transmitting. This can be problematic in dense deployments like shopping malls, stadiums, etc. Furthermore, and as we can observe in Figure 3-left, the device throughput can fluctuate significantly (>100Mbps in some cases). On the other hand, the aggregate throughput of 5G remains the same regardless of the number of devices (Table 2) and all devices achieve predictable throughput (Figure 3-right). This is due to the centralized radio resource scheduler, that ensures the network's reliability.

To evaluate the coverage of the two technologies, we run a continuous downlink UDP iperf test to one of the phones, as we walk around the building (floorplan in Figure 2). Our building has elevator shafts (the square in the middle), where we cannot measure coverage. We report an approximate coverage of Wi-Fi and 5G in Figures 2(a) and 2(b). We see that 5G approximately doubles the coverage, which illustrates the point that 5G can improve coverage in challenging enterprise settings with limited deployment options. Note that some reports, such as [11], claim up to 5× better coverage for 5G.

Benefits of open interfaces and network APIs – Next, we demonstrate how we can leverage the open 5G interfaces to optimize connectivity. We consider a challenging scenario (e.g., a shopping mall), where access points cannot be deployed uniformly for full coverage. We deploy one more radio unit, 5 meters away of the first one (see Figure 2(c)). We then build distributed MIMO as an add-on software service based on our architecture (Figure 1) and deploy using the two radios. The service taps into the open fronthaul interface that connects the RAN with the RUs (Figure 4). It modifies the destination addresses of the fronthaul packets between the server and the radios, and maps the packets of different antennas to different radios (illustrated with different colors in Figure 4). This turns the two RUs to a single distributed MIMO RU. As we see in Figure 2(c), this significantly improves the coverage and throughput to cover almost the entire floor, while keeping all radios in one corner of the building.

Other services can also be implemented in a similar way. For example, one could leverage network slicing APIs to provide QoS guarantees to applications in terms of throughput and latency (e.g., as demonstrated in works like [9, 12, 18]). Similarly, network APIs could also be used for cross-layer optimizations (e.g., video bitrate

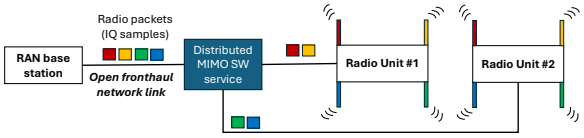
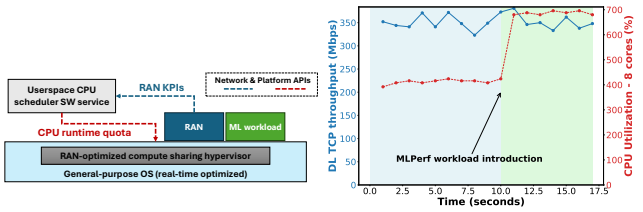


Figure 4: High-level overview of distributed MIMO design.



((a)) Design sketch. **((b)) Effects of CPU multiplexing.**

Figure 5: CPU compute sharing between RAN and ML workload.

adjustments based on signal quality [30]). Finally, such APIs could also be used for optimizing managmeent tasks, like for example RAN software upgrades with zero downtime [29].

Compute sharing – Here, we demonstrate the benefits of sharing the compute infrastructure between the cellular workload and the edge AI applications, to improve the utilization of the edge. For this example, we leverage a modified version of the RAN compute sharing technique proposed in [19]. As illustrated in Figure 5(a), we build an external userspace CPU scheduler as an add-on software service (Figure 1) that leverages platform and cellular APIs. Similar to [19], the scheduler collects real-time KPIs from the RAN and predicts its runtime for the upcoming period. It uses this prediction to assign the RAN process a CPU runtime quota using cgroups, allowing other workloads to run on the same CPU cores, while the RAN is idle. Using this mechanism, we multiplex the srsRAN workload with an ML workload (MLPerf benchmark for image recognition). Specifically, we deploy srsRAN on eight CPU cores of our server, which is the minimum number required to achieve full capacity. We then attach a UE, and we generate more than 300 Mbps iperf TCP traffic. As illustrated in Figure 5(b) (seconds 1-10), this results in an average CPU load of 400% across the eight CPU cores. We then deploy the MLPerf benchmark on the same eight CPU cores at second 10, which increases the CPU utilization to more than 85%, without affecting the throughput of the UE, effectively allowing us to reclaim approximately 3 CPU cores for the ML tasks.

As pointed out in [19], the latency of the workloads competing for the compute resources with the real-time CPUs of the RAN is expected to increase. However, we argue that those resources could be used to multiplex non-latency critical workloads, instead of real-time ones, freeing up CPU cores, to be used by the real-time industrial ML workloads. The shared workloads could include latency tolerant industrial edge ML-based workloads (e.g., Retrieval-Augmented Generation – RAG), or management workloads, like security scanners, orchestrators, software updates etc., which today require additional overhead CPU cores. It could also include the multiplexing of real time and non-real time threads of the RAN workload. It should be noted that the latency of the RAN workload itself will not be affected by this multiplexing, given that the RAN is treated as a high-priority workload (as already shown experimentally in [14]).

Cost – Finally, we discuss the cost dimension and how the advances discussed in Section 4 have drastically decreased it. We measure that we can fit eight 4×4 100MHz srsRAN cells on a server with 32 CPU cores with an Intel ACC100 accelerator. The approximate hardware cost for our edge deployment when dimensioned for eight cells is $25K, which includes the cost of the server, the radio units, a switch, a PTP grandmaster clock and the cabling. We note that this cost is at least an order of magnitude lower than the cost of deploying a telco-based solution of similar dimensions using licensed spectrum and conventional cellular technologies.

Next, we perform the same exercise for a Wi-Fi deployment of similar dimensions. We consider sixteen access points, to provide similar coverage to the eight 5G cells. Furthermore, we assume a server to run the edge applications, with similar capabilities as the one used for the cellular deployment, but with half the CPU cores, taking into account that approximately 50% of the CPU cores will be utilized by the RAN on average. We estimate that the cost of this deployment will be approximately $10K, which is 60% less compared to the cellular one. Based on this, we observe the following:

(1) The cost of the 5G RUs is about $5-10\times$ higher than WiFi. Considering that the RUs of both technologies are similar in terms of design (OFDMA-based), we argue that the cost difference is mainly due to the fact that there are currently very few private 5G deployments, driving the production costs up. For example, the Foxconn units that we used in our experiments are FPGA-based, rather than relying on cheaper ASICs, due to the lack of production at scale. We foresee that this cost will equalize, as the 5G deployment scale increases. This, combined with the fact that significantly less 5G radio units are required to provide the same level of coverage as WiFi (cf. [11]), would bring the costs of the two solutions to the same level.

(2) The 5G RAN stack has more complexity than WiFi, due to the large number of 3GPP features defined in the specifications, which increase the development and management cost of the RAN software. We argue that the majority of those features target telco networks and are not required for private 5G deployments in industrial scenarios. As such, one could have a much simpler (and thus cheaper) RAN software stack with basic features, which could be enhanced with add-on services on an as-needed basis, using open interfaces and APIs. As a concrete example, we consider the srsRAN stack that we used for our evaluation, which is open source, but provides great levels of performance and reliability. One could tune this stack with external applications (e.g., a network slicing scheduler) using the exposed APIs, to implement specific industrial use cases. We foresee that this approach will significantly drive down the cost of this part.

In addition to the above, the reliability of cellular deployments can also introduce long-term cost savings. For example, the real world case study of steel wheel manufacturing in [11] demonstrates that private 5G can be more cost efficient compared to WiFi, due to the significantly reduced unplanned downtime of the factory operations.

7 Concluding remarks and future work

Supported by the analysis of our case study, we reaffirm our belief on the benefits of a converged industrial edge that will cater the needs of reliable connectivity and intelligent edge applications. In our view,

this transition will be gradual, starting with simpler edge deployments that will offer basic connectivity and applications (e.g., edge data filtering/processing, simple control), followed by more sophisticated deployments with advanced capabilities, as more sophisticated AI applications come out and the cellular ecosystem matures.

We also believe that this transition creates an exciting opportunity for research in topics at the intersection of mobile, virtualization, hybrid cloud and applied AI. In line with this, we provide some example directions of future work:

Efficient compute sharing – The problem of efficient edge compute sharing between cellular and AI workloads is still at its infancy. Virtualization mechanisms offering strong isolation guarantees (e.g., VMs) incur overheads that can be significant for the resource constrained edge. On the other hand, new types of virtualization and isolation are also emerging (e.g., WebAssembly). Identifying the right mixture of virtualization technologies is an open research question.

Applications – The cellular network APIs create an exciting opportunity to innovate in the applications space. This includes identifying new ways with which the industrial edge applications can leverage the network to provide more advanced capabilities (e.g., via RF sensing, real-time control, etc.), as well as creating software services that can drive down the cost of the edge infrastructure, by leveraging ideas, like the distributed MIMO setup that was presented in this work.

Hybrid cloud – While autonomy is an important requirement for many industrial verticals, we believe that the cloud is still crucial, considering the abundance of resources that it offers. Figuring out the right way in which the cloud will interplay with the edge, in terms of orchestrating workloads across the compute fabric, while meeting their QoS requirements and minimizing the cost is an open problem.

Radio access technology convergence – While cellular connectivity might end up being the preferred technology for the industrial AI edge, we believe that Wi-Fi still has a major role to play. To become more competitive, Wi-Fi will have to provide similar capabilities to those of cellular in terms of expressive and open APIs. Identifying how those APIs will look like and figuring out their seamless integration with cellular is an open research question.

Acknowledgments

We thank the anonymous reviewers and our shepherd, Aruna Balasubramanian, for their invaluable feedback. This work was supported by the Open Networks Programme within the UK Department for Science, Innovation and Technology.

References

[1] [n. d.]. Airbus to replace Wi-Fi with 5G in "all industrial areas" within five years. https://www.rcrwireless.com/20241112/private-5g/airbus-private-5g. Accessed on 02.1.2025.

[2] [n. d.]. Engineering Insights: accelerating the 5G SA PHY layer processing with the ACC100. https://srs.io/engineering-insights-accelerating-the-5g-sa-phy-layer-processing-with-the-acc100/.

[3] [n. d.]. Tesla continues its 5G private network push in Shanghai. https://www.fierce-network.com/wireless/analyst-tesla-continues-its-5g-private-network-push-shanghai. Accessed on 02.1.2025.

[4] AI-RAN Alliance. 2023. Integrating AI/ML in Open-RAN: Overcoming Challenges and Seizing Opportunities. https://ai-ran.org/wp-content/uploads/2024/08/Integrating_AIML_in_Open-RAN_Overcoming_Challenges_and_Seizing_Opportunities-1.pdf. Accessed on 01.10.2024.

[5] Ganesh Ananthanarayanan et al. 2024. Distributed AI Platform for the 6G RAN. (October 2024). https://www.microsoft.com/en-us/research/publication/distributed-ai-platform-for-the-6g-ran/

[6] Eneko Artetxe et al. 2023. Wireless Technologies for Industry 4.0 Applications. *Energies* 16, 3 (2023). https://doi.org/10.3390/en16031349

[7] AT&T. 2023. ATT to Accelerate Open and Interoperable Radio Access Networks (RAN) in the United States through new collaboration with Ericsson. https://about.att.com/story/2023/commercial-scale-open-radio-access-network.html. Accessed on 01.10.2024.

[8] Paramvir Bahl et al. 2023. Accelerating Open RAN Research Through an Enterprise-scale 5G Testbed. In *Proceedings of ACM MobiCom '23*. 1–3.

[9] Arjun Balasingam et al. 2024. Application-Level Service Assurance with 5G RAN Slicing. In *USENIX NSDI '23*. 841–857.

[10] Sanjit Biswas et al. 2015. Large-scale Measurements of Wireless Network Behavior. *SIGCOMM Comput. Commun. Rev.* 45, 4 (Aug. 2015), 153–165. https://doi.org/10.1145/2829988.2787489

[11] Celona and Mobile Experts. 2022. Industrial Private Cellular Business Case. https://pages.celona.io/hubfs/Act-on%20Media%20Files/Celona%20Whitepaper%20-%20%20Industrial%20Private%20Cellular%20Business%20Case.pdf. Accessed on 01.10.2024.

[12] Yongzhou Chen et al. 2023. {Channel-Aware} 5g {RAN} slicing with customizable schedulers. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 1767–1782.

[13] Krishna Chintalapudi et al. 2012. WiFi-NC : WiFi Over Narrow Channels. In *USENIX NSDI '12*. USENIX Association, San Jose, CA, 43–56.

[14] Cambridge Consultants. [n. d.]. Wireless breakthrough for the Ocado Smart Platform. https://www.cambridgeconsultants.com/project/wireless-breakthrough-for-the-ocado-smart-platform/. Accessed on 01.10.2024.

[15] Ericsson. [n. d.]. Powering Industry 4.0: How the 5G ecosystem is helping drive next-gen industrial digitalization. https://www.ericsson.com/en/blog/2024/10/equipment-manufacturers-utilizing-5g-ecosystems. Accessed on 02.01.2025.

[16] Ericsson. 2024. 5G spectrum for local industrial networks. https://www.ericsson.com/en/reports-and-papers/white-papers/5g-spectrum-for-local-industrial-networks. Accessed on 10.10.2024.

[17] Xenofon Foukas et al. 2017. Network Slicing in 5G: Survey and Challenges. *IEEE Communications Magazine* 55, 5 (2017), 94–100.

[18] Xenofon Foukas et al. 2017. Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture. In *Proceedings of the 23rd annual international conference on mobile computing and networking*. 127–140.

[19] Xenofon Foukas and Bozidar Radunovic. 2021. Concordia: teaching the 5G vRAN to share compute. In *Proceedings of ACM SIGCOMM '21* (Virtual Event, USA). 580–596.

[20] Intel. 2023. 4th Gen Intel Xeon Scalable Processor (Codename: Sapphire Rapids Edge Enhanced). https://cdrdv2-public.intel.com/784461/784461_SapphireRapidsEE_SpecificationUpdate_002.pdf. Accessed on 01.10.2024.

[21] Intel. 2024. Intel Builds on vRAN Momentum with New AI Development Kit, Future Intel Xeon Processor. https://community.intel.com/t5/Blogs/Tech-Innovation/Edge-5G/Intel-Builds-on-vRAN-Momentum-with-New-AI-Development-Kit-Future/post/1572061. Accessed on 10.10.2024.

[22] Ruofeng Liu and Nakjung Choi. 2023. A First Look at Wi-Fi 6 in Action: Throughput, Latency, Energy Efficiency, and Security. *Proc. ACM Meas. Anal. Comput. Syst.* 7, 1, Article 25 (March 2023), 25 pages. https://doi.org/10.1145/3579451

[23] Shadi A Noghabi et al. 2020. The emerging landscape of edge computing. *GetMobile: Mobile Computing and Communications* 23, 4 (2020), 11–20.

[24] Emeka Obiodu. 2023. Pioneering 5G OpenRAN Advancements with Accelerated Computing and NVIDIA Aerial. https://developer.nvidia.com/blog/pioneering-5g-openran-advancements-with-accelerated-computing-and-nvidia-aerial/. Accessed on 01.10.2024.

[25] RCR Wireless News. 2024. Private 5G in Industry 4.0. Accessed on 01.10.2024.

[26] SoftBank. 2024. AI-RAN: Telecom Infrastructure for the Age of AI. https://www.softbank.jp/corp/set/data/technology/research/story-event/Whitepaper_Download_Location/pdf/SoftBank_AI_RAN_Whitepaper_December2024.pdf. Accessed on 02.1.2025.

[27] Verizon. [n. d.]. Verizon collaborates with NVIDIA to power AI workloads on 5G private networks with Mobile Edge Compute. https://www.verizon.com/about/news/verizon-nvidia-power-ai-workloads-5g-private-networks-mec. Accessed on 02.01.2025.

[28] Shikhar Verma et al. 2023. A survey on Multi-AP coordination approaches over emerging WLANs: Future directions and open challenges. *IEEE Communications Surveys & Tutorials* (2023).

[29] Jiarong Xing et al. 2023. Enabling Resilience in Virtualized RANs with Atlas. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[30] Dongzhu Xu et al. 2022. Tutti: coupling 5G RAN and mobile edge computing for latency-critical video analytics. In *Proceedings of ACM MobiCom '22*. New York, NY, USA, 729–742.

[31] Yuhan Zhou et al. 2024. AUGUR: Practical Mobile Multipath Transport Service for Low Tail Latency in Real-Time Streaming. In *USENIX NSDI '24*. USENIX Association, Santa Clara, CA, 1901–1916.