TECHNICAL OVERVIEW

AI-ON-5G

Delivering AI Applications at the Edge on a High-Performance 5G RAN

NVIDIA

In the decade leading up to 2030, technological developments and commercial use of AI and 5G will transform the enterprise landscape and accelerate economic growth. Combining AI at the edge and 5G under the NVIDIA AI-on-5G platform will turbocharge digital transformation and value creation. With industry-leading expertise in delivering enterprise AI on GPU-accelerated platforms, NVIDIA is well positioned for this new computing paradigm, offering enterprises, telecommunications operators, and cloud service providers (CSPs) the opportunity to integrate the 5G and edge AI ecosystem or convert the 5G gNB to an edge data center for AI workloads. NVIDIA AI-on-5G offers a converged, GPU-accelerated, software-defined computing platform, with a 5G virtual radio area network (vRAN) delivered as a software workload running side by side with other AI workloads. AI-on-5G at the edge opens new capabilities for smart cities, security systems, retail intelligence, industrial automation, and optimization of network capacity and utilization.

In this technical overview, we'll explore the drivers and opportunities for AI-on-5G and then describe the building blocks and technology strategy for the emerging AI-on-5G ecosystem.

EXECUTIVE SUMMARY

Artificial Intelligence is the most powerful technology force of our time. Enabled by 5G's high-performance, ultra-reliable, and highly secure connectivity, AI applied at the network edge will deliver connected intelligence across every industry, driving transformation of the enterprise landscape and accelerating economic growth. This opens up new use cases in smart cities, smart manufacturing, smart retail, automated warehouses, and more. The NVIDIA AI-on-5G platform is leading the industry push to combine AI and 5G at the network edge to accelerate this digital transformation that will create over $10 trillion in economic value.

Combining AI and 5G under NVIDIA's AI-on-5G platform unlocks new commercial opportunities for enterprises, telcos, and other stakeholders, whether on premises, in the field, or in the cloud. For enterprises, running AI applications at the edge on a high-performance 5G RAN is a key factor in realizing the Industry 4.0 vision of an interconnected and integrated system of cyber-physical systems. For telcos, deploying AI applications at the 5G edge creates new revenue sources, positioning AI to be the killer application for 5G and turning the discussion on 5G rollout from a capital expenditure discussion into a revenue and monetization discussion.

The AI-on-5G platform opens a new technical playbook. Today, 5G telecommunications infrastructure and the AI computing infrastructure are evaluated, deployed, and managed independently. This is an inefficient strategy, because both AI and 5G require computational oomph that can be provided by the same high-performance computing platform. Because of this, using the same computing platform to enable AI workloads to run seamlessly over a 5G network delivers both

technical and cost efficiencies. This brings lower total cost of ownership (TCO) for equipment, power, and space, enabling the auto-scaling and pooling of compute resources. It also delivers higher security for AI.

The NVIDIA AI-on-5G platform makes the combined AI and 5G opportunity a reality. Based on the NVIDIA EGX™ class of servers, it's a single, hyperconverged, GPU-accelerated, edge "data center in a box" that can deliver high performance for both AI workloads and 5G RAN. With the AI-on-5G platform, the 5G RAN is implemented as an additional software stack on the GPU-accelerated computing platform. This makes it possible for an enterprise to add and manage 5G connectivity as part of their IT infrastructure and a telco to transform all 5G gNBs into edge data centers. The five building blocks of the NVIDIA AI-on-5G platform are:

1. The EGX server
2. The NVIDIA Aerial™ 5G vRAN software development kit (SDK)
3. The Aerial A100 card
4. NVIDIA's portfolio of edge AI applications, including NVIDIA Metropolis™ and NVIDIA Isaac™
5. NVIDIA's partner ecosystem of equipment makers, independent software vendors (ISVs), developers, and more

INTRODUCTION

Across society, the next digital revolution will be shaped by connected intelligence at the edge. AI and 5G bring complementary capabilities to drive that digital transformation. AI is reshaping how enterprises across every industry segment, are doing their business. 5G offers superfast, wide-area, and ultra-reliable connectivity, enabling enterprises to link their AI systems, whether on premises, in the field, or in the cloud. Together, these two forces will revolutionize industrial sectors and create value in enterprise markets. IHS Markit expects the 5G-enabled value chain to generate $13.1 trillion of gross output by 2035.

> Embracing Industry 4.0

Since the early 1700s, human civilization has gone through three industrial revolution eras. Each revolution has been driven by transformative technologies that weren't present in the previous era.

INDUSTRY 1.0

With the advent of weaving looms in the nineteenth century, the first industrial revolution (Industry 1.0) was driven by mechanization.

INDUSTRY 2.0

Industry 2.0 kicked off at the start of the twentieth century as electrification replaced human and animal power in factories.

INDUSTRY 3.0

Toward the end of the twentieth century, Industry 3.0's computerization and digitalization introduced electronic controls and automation into industrial processes.

INDUSTRY 4.0

Today, the fourth industrial revolution is upon us, and we're on the cusp of another change: the creation of an interconnected and integrated network of cyber-physical systems. Industry 4.0 will rely on vastly improved computing power plus superfast and low-latency connectivity to blur the boundaries between the remote and the edge for cyber-physical systems.

Enterprises, telecoms operators, and CSPs are at different stages of deploying 5G and integrating it with enterprise AI applications. Today, 5G and AI infrastructures are evaluated, deployed, and managed separately. This is incredibly inefficient, as both AI and 5G require computational power that can be provided by the same platform.

As such, building them on the same computing platform and enabling AI to run seamlessly over 5G unlocks new use cases in enterprises—such as smart cities and automated warehouses—while delivering both technical and cost efficiencies. This reduces the TCO for valuable resources like equipment, power, and space and delivers higher security for AI.

NVIDIA AI-on-5G makes this opportunity a reality. Based on the EGX class of servers, it's a hyperconverged, GPU-accelerated, edge "data center in a box" that can deliver combined high-performance computing for both AI workloads and 5G RAN. With AI-on-5G, the 5G RAN is implemented as an additional software stack on the platform.

For enterprises, this means that the same computing platform for AI workloads can be used to support 5G RAN. For telcos, it offers the opportunity to transform every 5G base station to an edge data center, where the same computing platform can support 5G workloads and additional AI services. Bringing AI and 5G together under the AI-on-5G platform will reinforce linkages and turbocharge digital transformation.

MARKET DRIVERS: AI, EDGE CLOUD, AND 5G

> The Power of Data Center AI

The potential of AI is clear. AI applications—running on vast data centers in the cloud—have already transformed many industries. From retail to infotainment, AI has augmented human cognition and decision making, automated routine tasks, and unearthed unique insights from data. AI enables enterprises to detect fraud, improve customer relationships, optimize the supply chain, and deliver innovative products and services in an increasingly competitive marketplace.

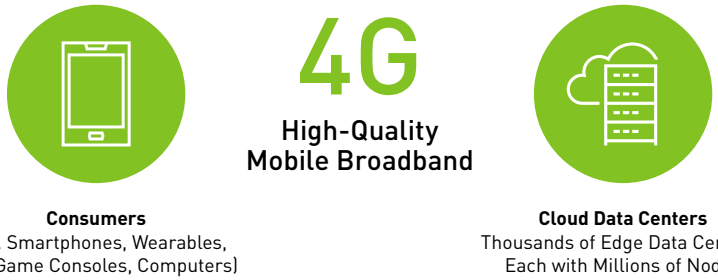


Figure 1: In the 4G era, consumers became confident about the availability and performance of their internet connectivity. This encouraged them to use on-demand services like cloud gaming and Netflix and Spotify, always-on services like Whatsapp and real-time services like Uber.

> Localization with Edge AI

Running AI applications at the edge brings the benefits of localization by processing data closer to where it's generated, captured, and used. This enables enterprises to react and adapt based on local conditions and requirements. Functionally, the edge cloud inherits most principles, mechanisms, and tools from the data center cloud. In addition, its proximity to end users brings new benefits such as lower latency for time-sensitive applications, transport and backhaul efficiencies for data-intensive applications, and data compliance for regulatory reasons. Crucially too, using the edge cloud limits the exposure of data packets to fewer network nodes, increasing reliability (less chance of equipment failure) and improving security (less network footprint that can be undermined).

Since AI is a highly compute-intensive process, having the right infrastructure optimized for the unique demands of AI workloads at the edge is essential. Data center clouds typically use lots of GPU-accelerated server infrastructure to ensure their high efficiency and effectiveness. Bringing this supercomputing capability to the edge requires integrating GPU-accelerated servers into the edge, together with specialized provisioning, management, and monitoring.

Operationally, the data center cloud and the edge cloud are a single continuum, with AI workloads moving along the spectrum based on what's most required. With this, enterprises can start off with data center AI and rebalance their workloads to edge AI as the edge infrastructure matures. However, before this can happen seamlessly, the connectivity between the data center and the edge needs to ensure a level of performance that has never been available on a mass-market level to enterprises.

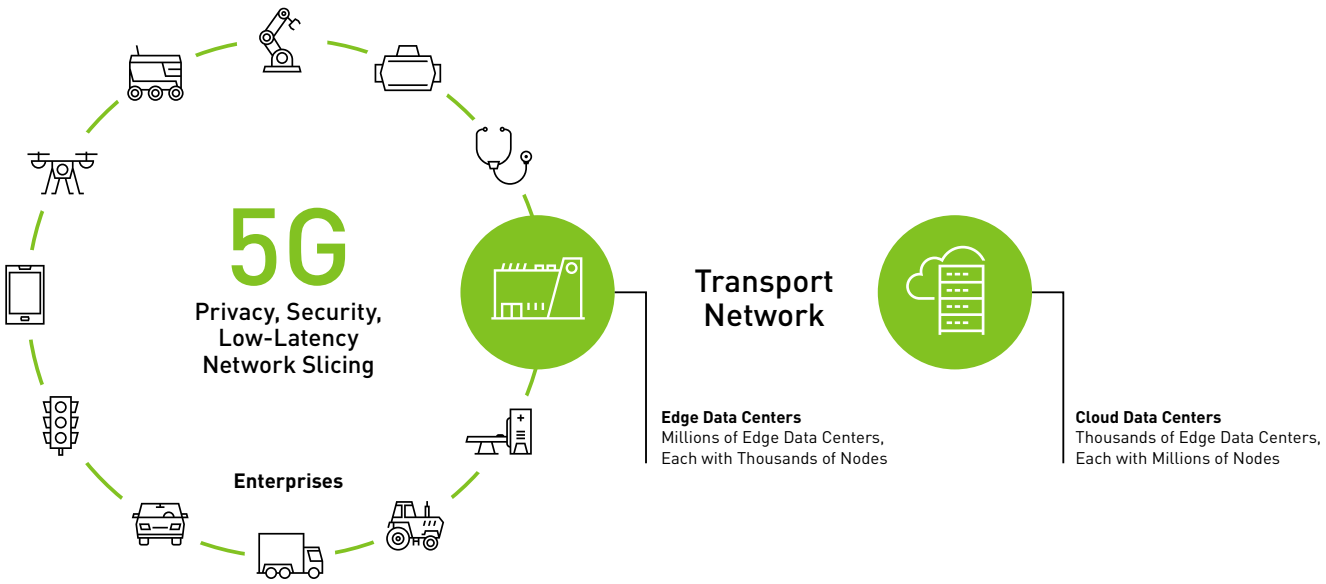


Figure 2: In the 5G era, enterprises will become confident about the availability, performance, and security of their internet connectivity. This will encourage them to deploy edge AI applications to monitor, operate, optimize, and manage their assets, whether they're on premises, in the field, or in the cloud.

> Hyperconnectivity with 5G

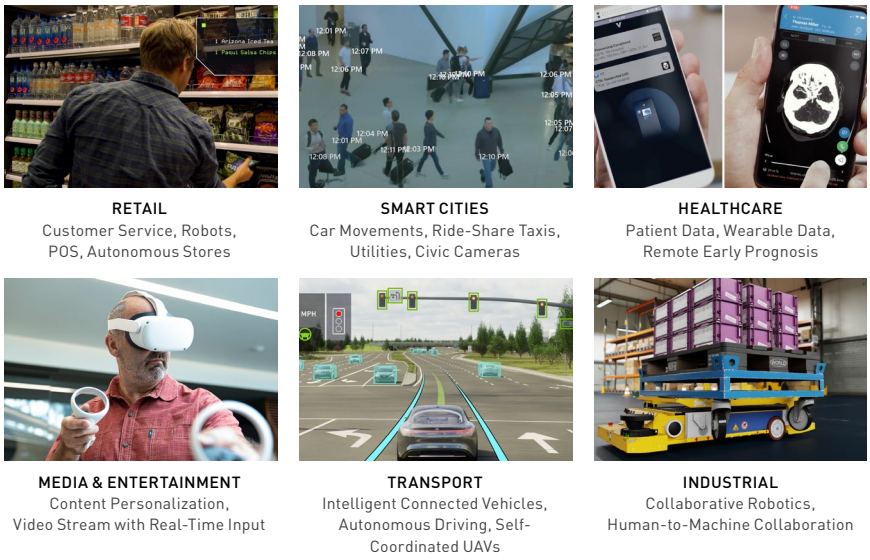
5G is poised to overcome the challenges of creating a seamless connection between data center AI and edge AI. Previously, it was a challenge to bring the benefits of data center AI to edge AI so that AI applications can run faster, avoid bandwidth bottlenecks, be better localized, and improve compliance with data residency rules. This is because these applications have often been constrained in coverage, mobility, latency, reliability, and security, by the interconnect medium, usually Wi-Fi or other proprietary options.

5G is addressing this challenge through traditional competencies combined with new capabilities. Primarily, 5G inherits cellular connectivity's wide-area coverage and seamless mobility, enabling enterprises to connect their equipment and run applications on on-premises, field-based, and cloud-based assets. In addition, the 3rd Generation Partnership Project's (3GPP) 5G standards (especially Release 16) is application-driven and positioned to address connectivity issues for machine-to-machine communication in enterprise and industrial scenarios. 5G offers low-latency, secure communications between devices, and it has the ability to communicate effectively and efficiently back from the edge and across the network to the big data centers everyone is familiar with.

Running AI applications on 5G will bring the true vision of AI applications anytime, anywhere to reality. It will bring extended coverage, mobility support, configurable quality of service, improved reliability, and enhanced security. This will enable, for example, more powerful deep learning training and inference—making AI better able to guide traffic flows, route autonomous vehicles, make factory robots more efficient at picking and packing goods, and much, much more.

OPPORTUNITY AND USE CASES

Connected intelligence, powered by AI and 5G, will create unprecedented economic value for all ecosystem stakeholders. PWC predicts that enterprise AI could contribute up to $15.7 trillion to the global economy by 2030, while IHS Markit expects the 5G-enabled value chain to generate $13.1 trillion of gross output by 2035. As the market for enterprise use cases reaches the tipping point, combining AI and 5G under the NVIDIA AI-on-5G platform has the potential to digitally transform every industry.



RETAIL
Customer Service, Robots, POS, Autonomous Stores

SMART CITIES
Car Movements, Ride-Share Taxis, Utilities, Civic Cameras

HEALTHCARE
Patient Data, Wearable Data, Remote Early Prognosis

MEDIA & ENTERTAINMENT
Content Personalization, Video Stream with Real-Time Input

TRANSPORT
Intelligent Connected Vehicles, Autonomous Driving, Self-Coordinated UAVs

INDUSTRIAL
Collaborative Robotics, Human-to-Machine Collaboration

Figure 3: Connected intelligence at the edge with 5G will transform every industry segment, impacting over $10 trillion in global economic value by 2035.

Enterprises, telecoms operators, and CSPs that deploy AI-on-5G will be able to handle both 5G and edge AI computing in a single, converged platform. This creates high-performance 5G RAN and AI applications to manage precision manufacturing robots, automated guided vehicles, drones, wireless cameras, self-checkout aisles, and hundreds of other transformational projects. It builds on NVIDIA's global leadership in AI and rich ecosystem of ISVs and partners that deliver applications across industries.

1. **Enterprises:** NVIDIA AI-on-5G is the optimal, cost-efficient approach for integrating 5G into enterprise operations. According to GSMA & ABI Research, over six million 5G cells will be deployed by 2027 to smart factories, fulfillment centers, and other enterprise, industrial, and public zones to provide localized connectivity solutions. Minimizing the capex burden of adding 5G to enterprise operations is the baseline for capturing value from 5G.
2. **Telecoms Providers:** Running AI applications on the 5G RAN is a big opportunity for telcos. As such, NVIDIA AI-on-5G provides a direct route to 5G monetization by making AI the killer app for 5G, shifting discussions about RAN investments to monetization opportunities. By 2027, the number of base stations in the world will more than double to 17.2 million. The capex burden of this investment and the uncertainties about new revenue opportunities are major discussion points for the industry and have made the 5G business case more a question of market readiness than profits.

THE TECHNOLOGY BUILDING BLOCKS OF AI-ON-5G

NVIDIA AI-on-5G is built on the NVIDIA EGX computing platform for enterprise AI. The roadmap is to go from today's colocated 5G + EGX edge cloud infrastructure to a future where the entire AI-on-5G

functionality is delivered on a single card. The platform is made up of these components:

1. EGX server
2. Aerial 5G SDK
3. Aerial A100 card
4. Edge AI applications
5. Partner ecosystem

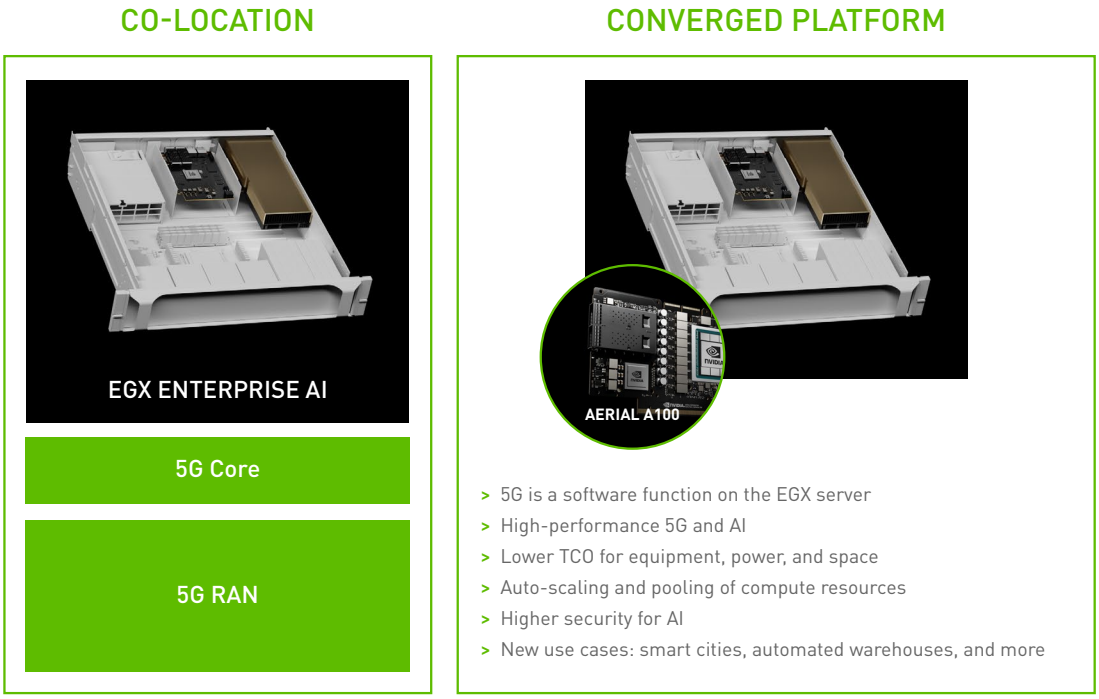


Figure 4: AI-on-5G roadmap. Today, 5G and edge cloud infrastructure are colocated, often with space, cost, and efficiency constraints. NVIDIA AI-on-5G will be delivered as a hyperconverged platform in a single Aerial A100 card.

THE NVIDIA EGX ENTERPRISE PLATFORM

The NVIDIA EGX Enterprise platform is a GPU-accelerated, cloud-native, end-to-end performance management, and software-defined infrastructure for data-intensive, graphics-rich, and computationally demanding enterprise applications. It consists of chips, systems, AI libraries, and a portfolio of applications from NVIDIA and ecosystem partners. NVIDIA EGX uses commercial-off-the-shelf (COTS) servers that are widely and commercially available from OEMs and ODMs and in the public cloud. It can be orchestrated to run mixed workloads utilizing the hardware assets of the GPU, CPU, and DPU, either virtually or as a container using Kubernetes on bare metal. Over 50 server partners from the world's top server vendors will be certified for NVIDIA EGX.

The NVIDIA EGX implementation with Aerial A100 is the first 5G base station that's also a cloud-native, secure, and hyperconverged AI edge data center capable of delivering the complete NVIDIA AI software suite.

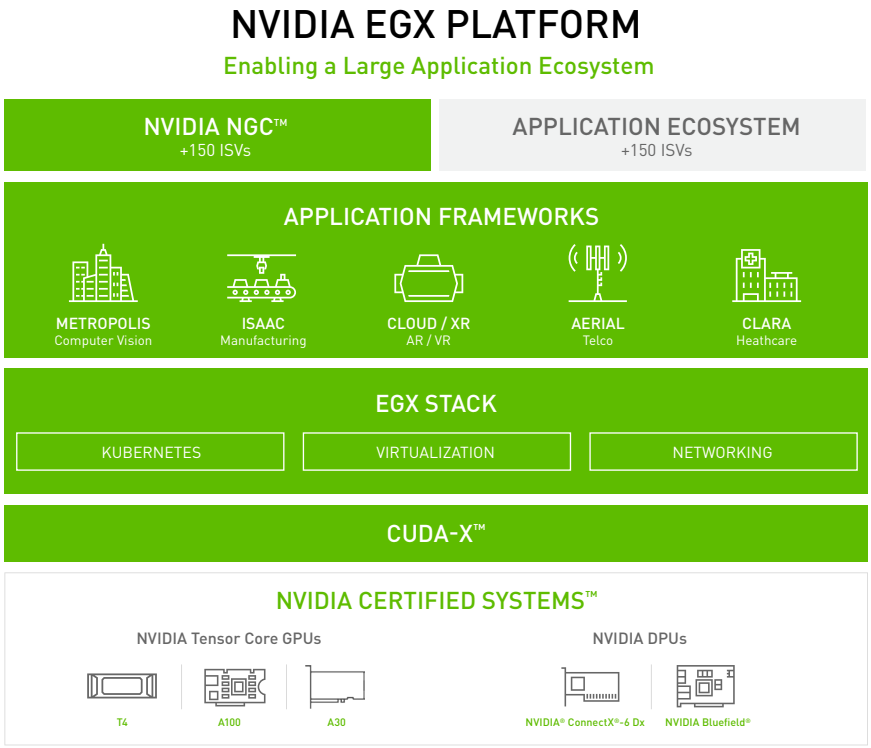


Figure 5: The EGX Platform. AI-on-5G is based on the EGX platform, bringing together NVIDIA-Certified Servers, an extensive software stack, and a portfolio of applications from a large developer ecosystem.

> NVIDIA Aerial 5G vRAN SDK

Aerial is a software development kit for building a 5G base station as an additional software stack on the EGX infrastructure. It empowers enterprises and telcos to support additional software platforms, allowing them to use the same computing infrastructure required for 5G networking to provide AI services, including enterprise, industrial, and consumer and residential workloads.

Aerial is a programmable, cloud-native, and scalable edge computing platform for 5G vRAN. Its CUDA Virtual Network Function (cuVNF) and CUDA Baseband (cuBB) SDKs provide highly programmable physical-layer signal processing and delivers extremely efficient Layer 1 (L1) processing, compared to the current alternatives. Its high-performance 5G RAN stack delivers network function acceleration capabilities, such as a 5G User Plane Function (UPF) core, AI-driven Near-Real-Time RAN Intelligent Controller (Near-RT-RIC), front-haul termination with ORAN-compliant timing, in-line security, and VNF offload, together with AI libraries and application SDKs. In addition, the Aerial SDK is highly scalable, with 100 percent COTS hardware support for cloud-native 5G applications.

Figure 6: Aerial 5G vRAN. The Aerial 5G vRAN stack is a set of SDKs that enable GPU-accelerated, software-defined 5G wireless RANs. Today, NVIDIA Aerial provides two critical SDKs: cuVNF and cuBB.

> NVIDIA Aerial A100 Card

The NVIDIA Aerial A100 card is a computing platform, designed for the edge, that combines AI and 5G into one EGX PCIe enterprise card. By adding a software-defined 5G RAN stack with full inline L1 acceleration running on the GPU, the EGX PCIe card transforms into a complete software-defined, ORAN-compliant, 7.2x split RAN compute base station. The NVIDIA Aerial A100 delivers up to 20 gigabits per second throughput and can process up to nine 100MHz massive multiple-input and multiple-output (MIMO) with 16 downlink/uplink and 8 up/downlink layers for 64T-64R radio.

The Aerial A100 contains three main elements:

1. The **GPU**, which runs the high-performance, full in-line L1 vRAN. The GPU is also capable of running the L2 scheduler and Near-RT-RIC algorithms, plus the AI multi-access edge computing (MEC) workloads.
2. The **BlueField DPU**, which combines NVIDIA network interface cards (NICs) and ARM® CPU computing to provide full infrastructure-on-chip programmability and high-performance networking. In the context of 5G, the DPU is also the termination point for the enhanced Common Public Radio Interface (eCPRI) ORAN-compliant front haul. Additionally, the programmable DPU fabric offloads the 5G UPF from the host CPU, accelerates network functions, and manages the software-defined networking. The advanced NVIDIA 5T-for-5G features on BlueField simplifies time

synchronization and data transmission across servers, GPUs, radios, and baseband units in wireless network rollouts, making 5G rollouts easier and more efficient. 5T-for-5G includes support for features such as precise time stamping, high clock accuracy for extremely precise timing accuracy (within 16 nanoseconds (ns)), eCPRI windowing, and time-bound packet steering.

3. The host **CPU**, which runs control plane functions, L2, and other functions.

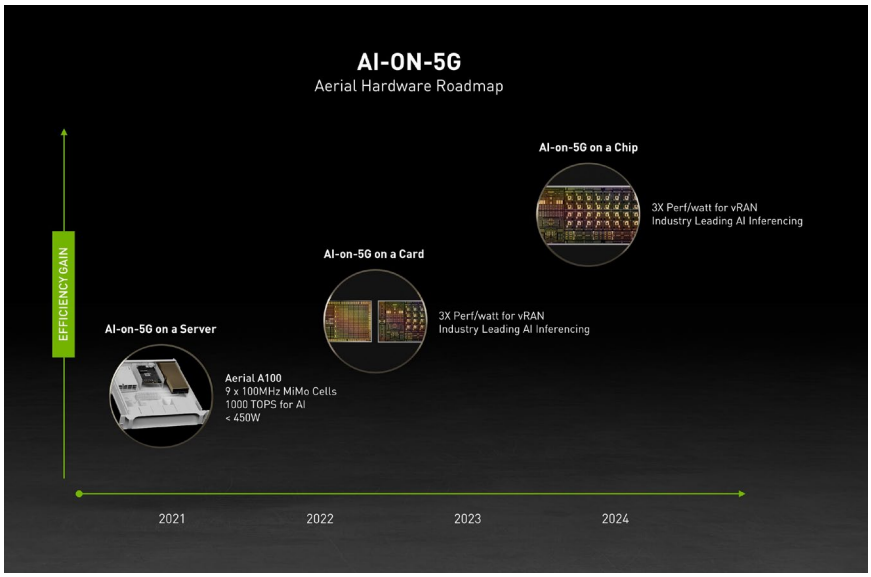


Figure 7: Aerial A100 hardware roadmap. The current hardware implementation is the Aerial A100 PCIe card attached to an x86 host. To further integrate the GPU, DPU, and CPU functions, the next release in 2022 will leverage two chips: the BlueField-3 (BF3) DPU and the NVIDIA Ampere architecture GPU. With the combination of a very high-performance ARM CPU cluster, next-generation NIC subsystem, and the highest-compute-capable GPU, BF3-Aerial A100 will deliver AI and high-performance vRAN on a single card.

> Edge AI Applications

NVIDIA has a large suite of edge AI applications that can be deployed on the AI-on-5G platform. In addition to Aerial, AI-on-5G is well suited for NVIDIA Metropolis™ (for computer vision), NVIDIA Isaac™ (for factory automation), NVIDIA Clara™ (for healthcare data analysis), and NVIDIA CloudXR™ (for immersive reality), plus many more. These SDKs are already supported by a large ecosystem of application developers, bringing deep expertise to the AI-on-5G ecosystem from day one.

The combination of Aerial and other SDKs unlocks the power of AI-on-5G. For example, a manufacturer may need to deploy private 5G networks to provide advanced connectivity and manage their plethora of IoT endpoints. Today, that manufacturer would need separate edge computers to process the resulting data and different ones for the 5G functionality. Likewise, a major city center may experience a situation where telcos want to deploy millimeter wave (mmWave) cells and city authorities need to deploy hundreds of cameras for traffic monitoring and security purposes. Today, these would be two

separate and independent hardware boxes, and any telco who seeks to support analytics for the cameras would need to deploy additional edge computing resources at the base station. With NVIDIA AI-on-5G on the EGX platform, these solutions can be delivered in a single computing infrastructure.

TECHNOLOGY STRATEGY

AI-on-5G provides a clear path to a new technology strategy for both the computing and telecommunications industries. In a way, a sort of musical-chair game is playing out between these two industries. While AI—running on vast data centers in the cloud and at the edge—is increasingly mainstream for the computing industry, it's still in its infancy in telecommunications. In tandem, the telecommunications industry is committing multi-billion dollars annually to deploying 5G networks, yet is unsure how to monetize their investments or how to tap into AI- and IoT-driven services.

In parallel, a paradigm shift is happening in the telecommunications industry. For the first time in history, cellular telecommunications technology is no longer being viewed as simply a means for people to talk or use the internet on their phones. Instead, 5G has been developed to additionally drive the digital transformation of industrial sectors and create value in enterprise markets.

In this emerging scenario, the burning question for 5G stakeholders is: Where and how do I leverage the multi-billion annual capex for 5G to tap into the AI-enabled opportunity? NVIDIA believes that the answer to this question is anchored in three core actions:

1. **Implement 5G as software and avoid asset duplication:** For enterprises seeking to add 5G to their operations, the option of software-defined 5G, running on the same computing platform as other AI workloads, is key to realizing their Industry 4.0 aspirations.
2. **Transform 5G base stations into AI-on-5G data centers:** The starting point of the new technology strategy is to convert 5G base stations to AI-on-5G data centers and to proposition existing edge data centers for 5G. This will extend the trend that's already evident with CSPs who offer 5G functionality on their platforms.
3. **Unlock revenue opportunities for 5G business cases:** With AI as the killer app for 5G, NVIDIA AI-on-5G offers a much more positive business case for 5G, turning it from a discussion on when to schedule capital investment into a discussion on how to create new revenue and value for an enterprise.

To learn more, visit: **www.nvidia.com/ai-on-5g**

NVIDIA