NVIDIA Aerial GPU Hosted AI-on-5G

Anupa Kelkar
NVIDIA
Santa Clara, USA
anupak@nvidia.com

Chris Dick
NVIDIA
Santa Clara, USA
cdick@nvidia.com

2021 IEEE 4th 5G World Forum (5GWF) | 978-1-6654-4308-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/5GWF52925.2021.00019

***Abstract*—In this paper we present the NVIDIA hyper-converged platform supporting 5G connectivity and Mobile Edge Computing (MEC). 5G connectivity is realized with our *Aerial* [1] GPU-based cloud native 5G gNB. We introduce *AI-on-5G* on a converged accelerator to showcase our innovation in being able to host Aerial vRAN baseband processing, AI/ML training and inference, data analytics and other workloads. In other words, a data center at the edge that is provisioned with 5G connectivity as a service. We describe 3 uses-cases that highlight how existing NVIDIA AI/ML development frameworks, together with Aerial, can be leveraged to bring Industry 4.0 to reality.**

As an open platform Aerial is positioned to be industry transformational by providing researchers with a platform for next generation wireless and AI research.

Aerial seeds the research ecosystem with a first-class out-of-the-box (OOB) experience with a standards compliant 5G NR PHY. Researchers can run the supplied 3GPP compliant test vectors, and perform over-the-air experiments, using standard servers equipped with a GPU-based PCIe card. The PHY code base can be tailored to support research that combines AI/ML with 5G wireless.

***Index Terms*—5G vRAN, machine learning, artificial intelligence, GPU, O-RAN, mobile edge computing (MEC), multi-service architectures, flexible and programmable RAN, robotics, industry 4.0**

I. Introduction

A variety of End-to-End (E2E) use cases will be satisfied by 5G wireless communication with its ability to deliver unprecedented data rates and low latency to an extraordinary amount of users, combined with AI services that are everywhere today. NVIDA GPUs will play a significant role in this evolution. Machine Learning (ML) and Artificial Intelligence (AI) today can be embedded for industry value creation. In manufacturing for example AI is being used to control machines, to predict issues in supply chain, detect defects and automate quality assessment. 5G – the fifth generation of wireless technology – represents a powerful catalyst for this digital transformation. It is multiplying the potential of Artificial Intelligence. These prospects are made possible by a huge leap forward in wireless communications in terms of spectral efficiency, throughput and reduced latency. We are at the cusp of combining two fundamental disruptive technology trends today that enable each other and whose fates are intertwined. AI is a key enabler but needs access to copious amounts of near real-time, quality data to be able to provide accurate predictions and results. 5G's massive device connections and fast speeds combined with edge computing power – in turn powers AI to "learn" and make smarter, faster and reliable decisions.

Aerial [1] *AI-on-5G* has been designed and built from the ground up on 3 foundational pillars. 1. Leverage. 2. Extend. 3. Innovation. Aerial leverages a rich and diverse ML and AI application ecosystem, application frameworks, and extends this by applying to specific use cases powered by 5G. With Aerial we extend the NVIDIA converged platform, stack and solution by offering connectivity as a service. GPU hardware platforms and converged stack will continue to bring new innovations in the field of signal processing and E2E services to enable evolution longer term towards 6G using AI/ML pipelines.

The outline of the paper is as follows. In Section II we provide an overview of the Aerial gNB architecture including the layer-1 PHY pipeline and discuss how data is efficiently moved from the eCPRI fronthaul interface to GPU memory for processing without involving the host CPU. Section III describes the Aerial over-the-air testbed. Section IV describes the concept of a hyperconverged platform and describes three use-cases employing Aerial for XR, computer vision and robotics. In Section V we highlight some of the future directions for Aerial. Finally, in Section VI we provide our conclusions.

II. Aerial System Architecture

Fig. 1 provides an overview of NVIDIA's wireless technology stack. The silicon foundation comprises a host CPU, NVIDIA GPU for signal processing, AI/ML compute, data analytics and other workloads, and an NVIDA Mellanox NIC for fronthaul connectivity. The next layer in the stack are the Aerial SDKs comprising of a set of libraries and APIs that provide GPU in-line accelerated PHY layer functionality in an ORAN 7.2 split configuration. Today there are two SDKs – cuBB [5] and cuVNF [5]. The top layer of the stack is layer-2 which today runs on a CPU.

The NVIDIA cuBB SDK [5] provides a GPU-accelerated 5G signal processing pipeline, including cuPHY [5] for L1 5G NR PHY signal processing. cuBB delivers unprecedented throughput and efficiency in a pure software flow by keeping all physical layer processing within the GPU's high-performance memory and exploiting the massive compute parallelism of the array of streaming multiprocessors (SMs) [6].

cuVNF is a GPU enabled Data Plane Development Kit (DPDK) that supports SyncE, meets the ITU G.8273.2 standard for Slave / GrandMaster boundary clocks, and enables peer-to-peer data movement between the NIC and GPU, so avoiding the need for CPU *memcopy* operations.

978-1-6654-4308-1/21/$31.00 ©2021 IEEE
DOI 10.1109/5GWF52925.2021.00019

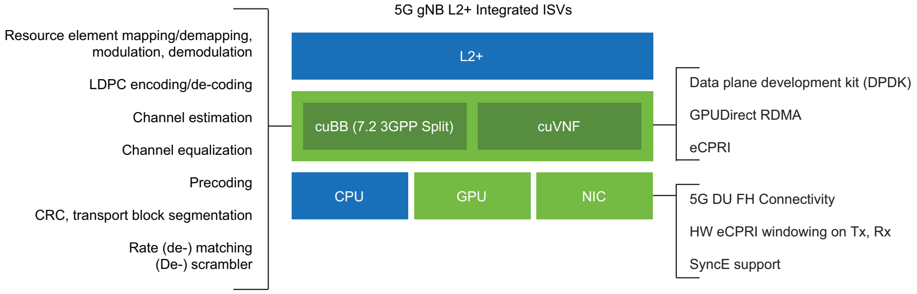


Fig. 1. The Aerial gNB stack comprising L2, physical layer (cuBB) and FH interface (cuVNF).

Fig. 2 provides a more detailed view of the complete Aerial gNB architecture, including some of the details of the physical layer (cuPHY) signal processing pipeline. All of the functions from transport block partitioning, channel coding, modulation and spatial processing are implemented in software as CUDA (Common Unified Device Architecture) [2] kernels. Even the LDPC decoder, the most computationally intensive function in the pipeline, is implemented entirely in software with no need for FPGA (Field Programmable Gate Array) or ASIC-style acceleration blocks. A completely softwarized approach to implementing the baseband processing reduces development time and provides flexibility compared to FPGA and ASIC implementations. A pure software approach is also the seque to future systems that will use AI/ML to optimize the base station and network via continuous learning and adaption to the environment.

A. Fronthaul Interface and Data Movement

One of the most critical components of any streaming signal processing application is the efficient management of data movement between the system resources. In this case, a high-performance eCPRI fronthaul (FH) interface between between the O-RAN Radio Unit (O-RU) and the GPU baseband processor is required to support the high throughput requirements of 5G NR wideband carriers combined with massive MIMO (mMIMO) antenna arrays. The FH interface is terminated using a Network Interface Controller (NIC) using an NVIDIA Mellanox Data Processing Unit (DPU) [8]. One model for moving data between GPU memory and the FH interface would be to employ the host system CPU. The CPU would move data from the NIC to system memory and use a *memcopy* operation to move the data from system memory to GPU memory. This is obviously inefficient. However, there is a much more efficient method, *GPUDirect Remote Direct Memory Access (RDMA)* [10], to move data between NIC and GPU memory. Fig. 3 illustrates the fundamentals of GPUDirect RDMA. Packetized data is presented to the NIC on the FH interface where it is split in to header and payload components. The payload data is delivered to GPU memory via the PCIe switch shown in the figure. The header information traverses the same PCIe switch and is written to host system memory. The packet header data structure in system memory also retains pointers to the payload data residing in GPU memory. Once a batch of payload frames has been delivered to GPU memory the GPU processing kernel can be launched.

III. Aerial Over-the-Air Testbed

Fig. 4 shows the Aerial wireless testbed. The NVIDIA wireless development kit ('DevKit') is shown to the right of the figure. The central section of the figure shows grand master clock generation and on the left the MIMO antenna panel communicating to a OnePlus handset is shown.

The devkit can be used as the foundation of a research platform for both 5G and 6G. The L1 CUDA kernels can be tailored to support PHY research. AI/ML functionality can be supported on one or both of the servers in the kit.

IV. Hyper-converged AI-on-5G Architecture

With industry-leading expertise in delivering enterprise AI on GPU-accelerated platforms, NVIDIA is already positioned to offer hardware and software innovation to enterprises, telecommunications operators, and cloud service providers (CSPs), including the opportunity to integrate 5G and the edge AI ecosystem for an edge data center. NVIDIA *AI-on-5G* offers a converged, GPU-accelerated, software-defined computing platform, with a 5G virtual radio area network (vRAN) 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 using a common infrastructure to deliver optimizations, efficiency, and lower overall cost of ownership.

Traditionally this infrastructure in a data center has been purpose built. As infrastructure needs evolve and as new services are deployed additional servers have been typically added to the mix. By leveraging the same NVIDIA EGX platform 6 that can enable the orchestration of multiple workloads, enterprises can reduce infrastructure costs, make systems functionally secure, and simplify system management. Reducing the number of components enables organizations to streamline operations, improve productivity, and reduce cost and complexity. Enterprises can lower both capital expenditure (CAPEX) and operating expenses (OPEX) by reducing the number of unique devices that must be kept on hand for maintenance and cutting the related costs of training and

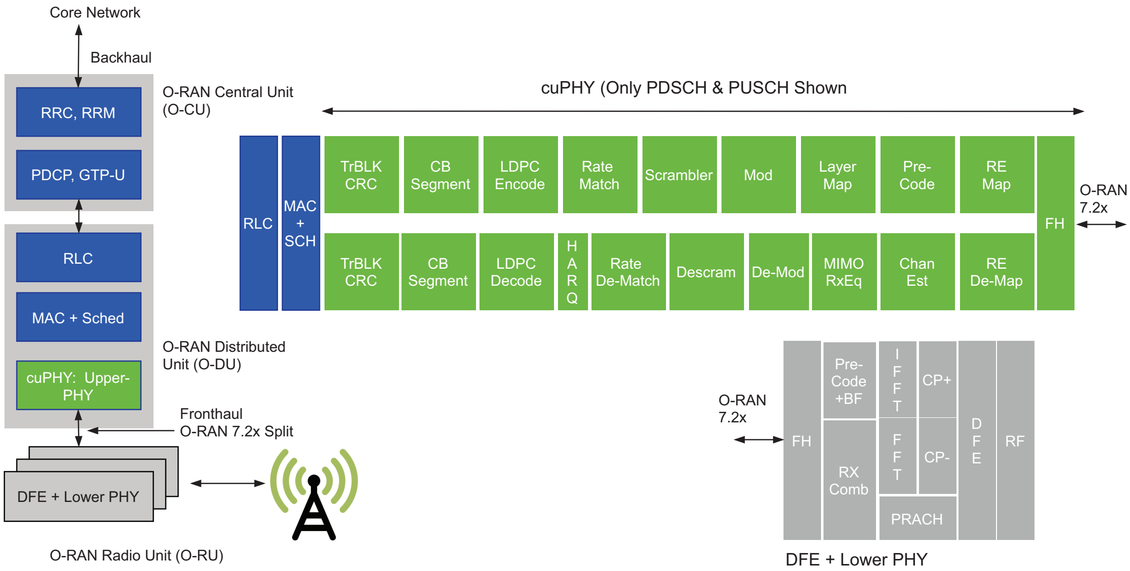


Fig. 2. Aerial system architecture. The green colored blocks are implemented as CUDA programs running on the GPU while the blue functions are supported on the host CPU.

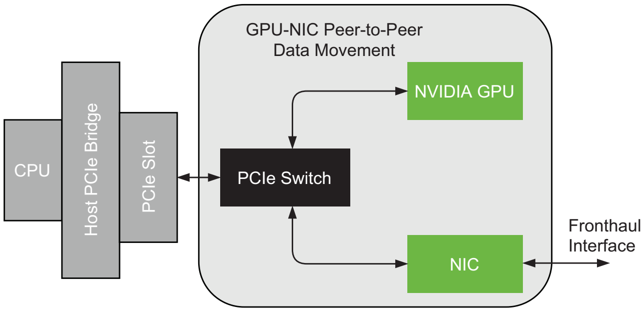


Fig. 3. NIC-GPU peer-to-peer data movement using GPUDirect RDMA. Fronthaul data is efficiently moved between the NIC and GPU memory without involving the host system CPU.

supporting staff. Reduction in per server power consumption is another plus of reduced number of servers through workload consolidation. Aerial DevKit v1.0 is capable of hosting vRAN, 5G CN (Core Network) and sample Edge AI services.

To re-iterate, consolidating workloads onto NVIDIA EGX with Aerial 5G connectivity offers tangible benefits through savings in maintenance, reduced training and support and simplifies integration. Instead of having a dedicated server or servers for each point solution, a single, GPU hosted High Performance Compute platform can host several services. By reducing the number of servers used by point solutions, we can expect considerable reduction in the the number of servers managed, which means less operational work. The attack surface of the network is minimized by reducing hardware, firmware, and software variants. Through open, inter-operable, non-purpose built, and multi-workload capable solutions system maintenance is simplified.

NVIDIA's BlueField [8] architecture combines an Ampere class GPU with ARM CPUs and Mellanox NIC. This converged accelerator will be used for NVIDIA's future 5G solutions and offers the benefits of tighter integration, latency reduction, power efficiency and ease of deployment. It also ensures optimal in-line 5G processing with GPUDirect RDMA , built-in network interfaces and functions and security offload in addition to hosting mixed workloads including vRAN and AI.

In a converged system the goal is to run multiple workloads on the one platform. GPU virtualization enables one physical GPU to be shared in a secure manner between multiple workloads. This capability is realized using *Multi-Instance GPU (MIG)* [11]. Several MIG partitions could be allocated to vRAN processing while the remaining instances are running machine learning workloads supporting, for example, a Smart City [12] application. The Smart City application would be easily designed using the Metropolis SDK [13] which provides application developers with video analytics libraries and a transfer learning toolkit to accelerate the development of custom ML models. A MIG partition could also be used to run online training to assist signal processing functions in L1 and the L2 scheduler for example.

AI-on-5G connects robots, smart devices and people during all phases of industrial automation design, deployment, operations and maintenance.

In the section below 3 use-cases that use Aerial for transport and NVIDIA CloudXR, Metropolis and Isaac for services to showcase high throughput, low latency and reliability are presented. Fig. 7 highlights the overall converged stack with NVIDIA EGX reference platform, application frameworks that use Aerial for 5G connectivity.

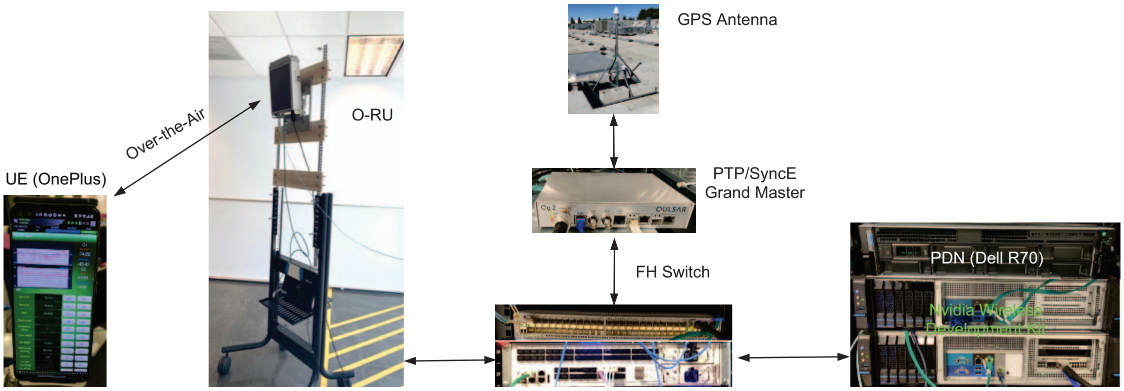


Fig. 4. Aerial 5G NR Over-the-Air MIMO testbed using the NVIDA wireless development kit.

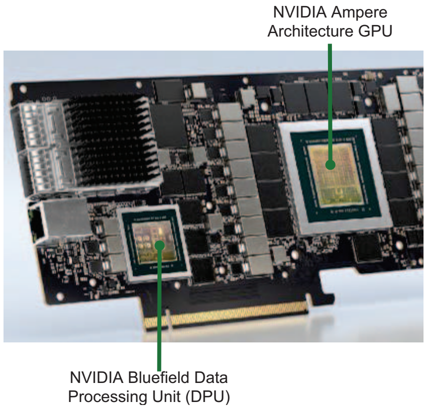


Fig. 5. Converged accelerator card showing the NVIDIA Ampere class GPU and *Bluefield* DPU (Data Processing Unit) [9].

1) Referring to Fig. 7, use-case 1 showcases design collaboration using NVIDIA CloudXR. In the automotive manufacturing cycle, with designers in different locations CloudXR allows design collaboration in real time. VR and AR provide immersive visualization for the designer and collaborators for quick design turnaround.
2) Use-case 2 showcases a defect detection use case during car manufacturing. The NVIDIA *Intelligent Video Analytics* framework allows visual inspection that develops a deep learning-based algorithm and trains it with examples of defects. Using the cognitive capabilities of visual inspection solutions quality issues can be resolved inline during manufacturing.
3) Use-case 3 demonstrates maintenance of a robot in distress in a car manufacturing plant. While robots nominally operate as automatically guided vehicles (AGV) and execute assigned payloads, there can be unexpected factory floor maintenance needed. The use case demonstrates how to accomplish remote tele-operations through the robot fleet mission controller using a digital twin and a robot mounted live camera feed. The tele-operator, with a virtual rendering of the physical plant, live orientation and positioning telemetry is able to operate the robot, guide it out of distress and re-enable its AGV operations.

V. Future Directions

While Aerial 5G is 3GPP Release-15 compliant today, enablement of industrial use cases will include support for Release-16 features that help improve performance, efficiency, provide better coverage, increase capacity, reduce latency and improve power efficiency. Ultra low latency and high reliability provides an opportunity for 5G to become the de facto connectivity solution for Industry 4.0.

Research on the use of AI/ML in the wireless domain is in its nascent phase right now. There is a lot of publishing, but for the most part the commercialization of AI/ML is yet to come. But it is clear, given the complexity of 5G and 6G networks, in the future, AI/ML will play a major role - many facets of 6G are already being planned with AI/ML as foundational technologies. AI/ML will ultimately be applied at all network layers, from the PHY through to network configuration and operation. Continuous learning, including online learning, will become the norm. Machine learning systems that are continually training and adapting to changing environmental (e.g. channel) conditions and other operational parameters will be at the core of next generation 5G/6G networks. The silicon technology that is critical to materialize these systems will need to support wireless signal processing in addition to having first class support for machine learning training and inference processing. NVIDIA GPUs, supporting massively parallel signal processing on streaming multiprocessors (SMs) and training/inference workloads on both SMs and tensor cores are uniquely positioned to address these requirements.

Equally importantly, and probably even more important, is the device programming model and availability of AI/ML development frameworks. The CUDA application programming interface (API) provides developers with a high-level abstraction for programming GPUs using a C/C++ syntax. The

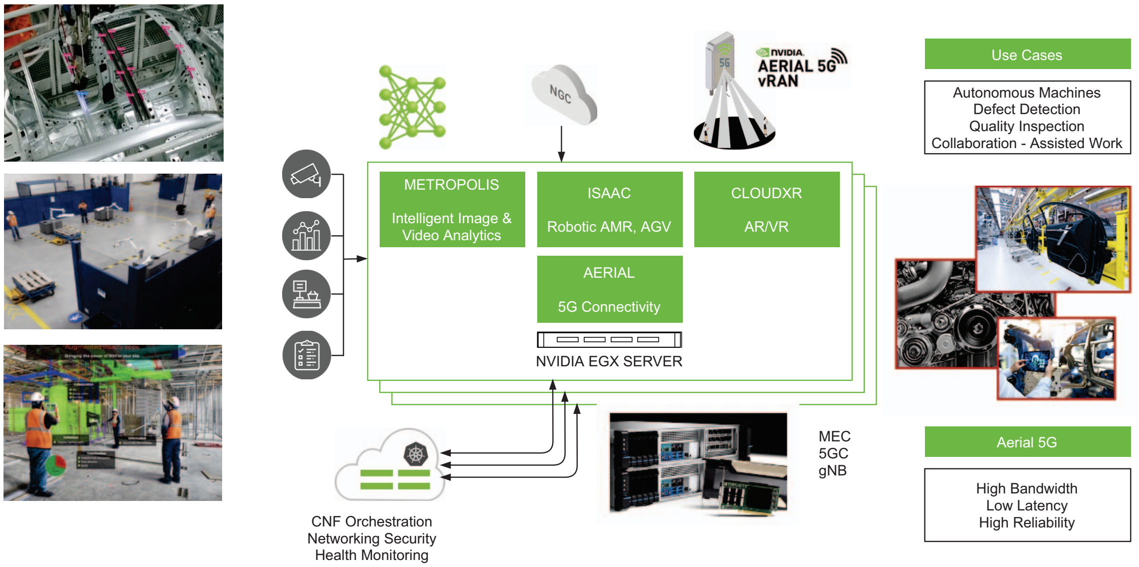
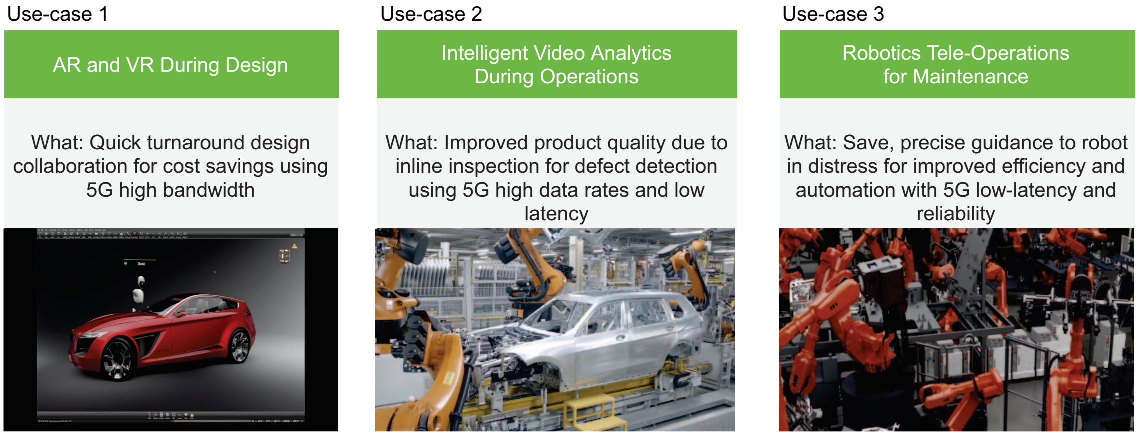


Fig. 6. Aerial EGX reference with AI-on-5G use-cases.



Use-case 1

AR and VR During Design

What: Quick turnaround design collaboration for cost savings using 5G high bandwidth

Use-case 2

Intelligent Video Analytics During Operations

What: Improved product quality due to inline inspection for defect detection using 5G high data rates and low latency

Use-case 3

Robotics Tele-Operations for Maintenance

What: Save, precise guidance to robot in distress for improved efficiency and automation with 5G low-latency and reliability

Fig. 7. NVIDIA AI-on-5G connecting robots, devices and people. The combination of AI and 5G enable all phases of industrial automation from design through to deployment, operations and maintenance.

large number of NVIDIA AI/ML development frameworks provide the means for researchers to rapidly test new concepts and for production teams to reach deployment. The confluence of GPU hardware, CUDA programming model combined with a rich set of AI/ML SDKs is unique in the industry and will power the future of wireless systems.

Incorporating digital twin in our AI-on-5G solution stack allow for continuous prototyping, testing, assuring and self-optimization of the living network. This technology is already being used across NVIDIA through our Omniverse and simulation frameworks for other enterprise applications. Depending on the specific use case there maybe unique requirements to consider. For example factories have a complex set of communication requirements (i.e., ultra-low latency, ultra-high reliability, high device synchronicity, time-sensitive networking, ultra-high levels of security and data privacy) and environmental challenges (i.e., surfaces impacting radio propagation, interference, zonal privacy, multi-tenancy) - Digital Twin gives private network operators a cost-effective mechanism to continuously model, plan, optimize, and assure how 5G connectivity benefits the smart factory.

VI. Conclusion

In this paper we have described a new and completely software-based approach to realizing a 5G NR gNB base station.

Historically, it has been challenging to meet the performance, cost and power requirements of a cellular base station with a purely software defined radio (SDR) platform. Typically ASIC or FPGA accelerators have been used to support the compute-heavy workload of, for example, channel coding - LDPC and Polar codes in the case of 5G NR. This in turn compromises the implementation paradigm of a SDR methodology. Aerial requires no such accelerators. The compute capacity of recent generation GPUs like Ampere [6] now makes an SDR approach for 5G base stations a reality. But it is more than just the silicon alone that is important. The CUDA programming model puts the massive compute capability of the GPU within easy reach of application programmers.

As the 5G rollout continues, and 6G research begins, the vRAN will continue to be important, but the role of artificial intelligence and machine learning will become increasingly prominent. As highlighted in Section IV, the combination of the GPU, CUDA programming model and AI/ML SDKs provides a unique hyper-converged platform that can support not only the vRAN, but also AI/ML inference, training, data analytic and other Multi Access Edge (MEC) workloads.

References

[1] NVIDIA, "NVIDIA Aerial," July 2021, https://developer.nvidia.com/aerial-sdk. [Accessed: 12 July 2021].

[2] NVIDIA, "CUDA Toolkit," July 2021, https://developer.nvidia.com/cuda-toolkit. [Accessed: 12 July 2021].

[3] NVIDIA, "The NVIDIA EGX enterprise platform," July 2021, https://www.nvidia.com/en-us/data-center/products/egx/. [Accessed: 12 July 2021].

[4] Veronica Quintuna Rodriguez, Fabrice Guillemin, Alexandre Ferrieux and Laurent Thomas, "Cloud-RAN functional split for an efficient fronthaul network," 11 Jan. 2021, https://arxiv.org/pdf/2101.04216.pdf, [Accessed: 12 July 2021].

[5] NVIDIA, "NVIDIA Aerial," https://developer.nvidia.com/aerial-sdk July 2021, [Accessed: 12 July 2021].

[6] NVIDIA, "Ampere GA102 GPU Architecture," July 2021, https://www.nvidia.com/content/PDF/nvidia-ampere-ga-102-gpu-architecture-whitepaper-v2.pdf, [Accessed: 12 July 2021].

[7] NVIDA, "ConnectX-6 DX," July 2021, https://www.nvidia.com/en-us/networking/ethernet/connectx-6-dx/ [Accessed: 12 July 2021].

[8] NVIDA, "NVIDIA Bluefield Data Processing Units," July 2021, https://www.nvidia.com/en-us/networking/products/data-processing-unit/, [Accessed: 12 July 2021].

[9] NVIDA, "NVIDIA Converged Accelerators," https://www.nvidia.com/en-us/data-center/products/egx-converged-accelerator/ [Accessed: 12 July 2021].

[10] NVIDA, "NVIDIA GPUDirect,", July 2021, https://developer.nvidia.com/gpudirect, [Accessed: 12 July 2021].

[11] NVIDIA, "Multi-Instance GPU user guide," July 2021, https://docs.nvidia.com/datacenter/tesla/mig-user-guide/index.html#abstract, [Accessed: 12 July 2021].

[12] NVIDIA, "Smarter cities through AI, " https://www.nvidia.com/en-us/industries/smart-cities/, [Accessed: 12 July 2021].

[13] NVIDIA, "NVIDIA Metropolis," July 2021, https://www.nvidia.com/en-us/autonomous-machines/intelligent-video-analytics-platform/,[Accessed: 12 July 2021].