NVIDIA AI Aerial: FAQ

Based on Q&A from the NVIDIA 6G Developer Day held on 15 October 2024

| December 2024

Document History

Doc_Number

Version	Date	Authors	Description of Change
01	December, 1, 2024	Emeka Obiodu	Initial release

Table of Contents

Section 1: Aerial CUDA-Accelerated RAN

Aerial CUDA-Accelerated RAN		
	Questions	**Answers**
1.1	What are the Aerial components in the 5G stack?	In Aerial, L1 signal processing implemented in the cuPHY library is accelerated in the GPU and control plane in the CPU. We also provide GPU acceleration to the MAC scheduler with cuMAC library and other parts of L2 can execute on the CPU.
1.2	How does a developer get started with the Aerial software stack?	All Aerial software stacks are available through membership of the NVIDIA 6G Developer Program.
1.3	Who manages workload distribution/selection between the CPU, GPU, and NIC? Is there a scheduler in CUDA?	The application software would be responsible for managing the workload distribution across CPU, GPU and NIC.

1.4	Does the Aerial platform support other waveforms such as 4G and DVB-S2? Or what would be required to integrate those waveforms onto this platform?	At this time only 5G waveforms are supported. The 4G transmit and receive chains are not implemented in Aerial. Since the receiver is fully software defined, in general, any signal processing receiver can be implemented.
1.5	Is there an Aerial specific version of CUDA API?	No. All CUDA APIs used in Aerial are available with the installation of CUDA toolkit.
1.6	Are the CUDA developer tools accessible and free to use?	Yes, the developer tools are available as part of the CUDA toolkit. Please see https://developer.nvidia.com/cuda-zone
1.7	Is the infrastructure on VMs or containers?	Aerial supports containerization.
1.8	Is the flow of I/Q samples of FH to cuBB inline?	Yes, using GPUDirectRDMA the NIC is able to direct DMA into and out of GPU memory.
1.9	In Aerial CUDA-Accelerated RAN, does loading vary from slot to slot and how do you prevent the processing from exceeding the time budget?	The application software has to impose the checks to prevent processing load from exceeding the GPU capacity. This can be done by polling GPU work using cudaEventQuery.
1.10	In three cell setup, can each cell have different numerology or bandwidth configuration?	The cells can have different bandwidth configurations. It is recommended that cells of the same numerology are grouped together since they would have the same processing budgets.
1.11	Are we able to dynamically manage GPUs clusters allocated between RAN and AI? If yes, how?	MIG (Multi-instance GPU) partitioning is not dynamic. It can only be done when the GPU is idle.
1.12	Can kernels that are provided through a library such as cuBLAS also be batched together or organized into a graph?	The CUDA stream capture interface may be used to include library calls while creating a graph. Please see https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#creating-a-graph-using-stream-capture

1.13	Are DMAs (Dynamic Memory Access) used for hiding memory access latency?	DMA can be used to overlap copies with compute. This way the compute engines can be kept busy doing useful work while the copies occur in the background.
1.14	How is the batch size in Aerial CUDA-Accelerated RAN determined?	The batch size needs to be determined empirically. There are several factors which influence the batch size computation: Workload configuration, resource footprint of the implementation, processing latency budgets, etc.
1.15	What is the SIMD/vector processing capability of say A100/H100 GPUs?	Each A100/H100 SM (streaming multiprocessor) is 4 way super-scalar and thus can execute 4 concurrent warps (32 threads). The threads in the warp are SIMT (Single Instruction Multiple Threads) can execute in a SIMD fashion or independently. On H100 there are 132 such SMs.
1.16	As security becomes increasingly critical in AI-driven RAN solutions, how does NVIDIA view the role of quantum-enhanced cryptographic frameworks, such as Skywise AI's QuantumGuard+, in safeguarding next-gen telecom infrastructure?	The termination point for encryption is at PDCP, so the DU and RU are not able to access (i.e. decrypt) the user plane and control plane. Having said that, Aerial only provides a reference implementation. gNB vendors can choose to adopt additional frameworks as they see fit to meet their business goals.
1.17	Does pyAerial use torch/Sionna(TensorFlow) for AI components and cuPHY for the conventional components?	pyAerial currently makes use of cuPHY only; in release 1.2, it will support JIT python and we will add Torch support for AI.
1.18	Does the GPU and CPU share the same system clock? If not, how do you sync between L2 processed in CPU and L1 in GPU?	The CPU and GPU do not share the same system clock.
1.19	How is QoS/Slicing handled in Aerial CUDA-Accelerated RAN?	GPU sharing is handled via two mechanisms MPS and MIG.

1.20	Could you please elaborate on the use of graphs, and explain how they are applied in comparison to the legacy RADIO context landscape (topology, configuration , alarms, FCUPS, etc.)?	The example usage of graphs in the physical layer is presented in slides 14, 15. Need clarification on "legacy RADIO context landscape" (describing specific use cases might help) to be able to comment on how.
1.21	What is the maximum number of thread blocks supported in a cluster?	The cudaOccupancyMaxActiveClusters can be used to query the number of supported clusters for a given kernel. In Hopper the max cluster size is 8 with an option of 16. Please see https://docs.nvidia.com/cuda/hopper-tuning-guide/index.html#thread-block-clusters
1.22	How many stream priorities exist?	The number of CUDA stream priorities can be queried using cudaDeviceGetStreamPriorityRange API. As of this writing there are 6 stream priorities.
1.23	Can the Host-to-Device and Device-to-Host copies overlap?	Yes
1.24	Does low latency adversely affect QoS? If so, what are the tools that would help to collect metrics that support high QoS?	Yes, latency can be impacted by QoS. GPU sharing mechanisms MPS and MIG are ways to support QoS. GPU kernel latencies can be measured with CUPTI.
1.25	How much of these are automatically optimized by CUDA without having to explicitly program them?	Just like any C/C++ program, the CUDA program does not optimize by itself outside of the optimizations from the compiler. We recommend profiling the code using NsightSystems and NsightCompute to identify and optimize bottlenecks. If a program written in CUDA needs optimization we recommend following the best practices in https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

1.26	In vectorization what is the max length supported? Similar to AVX 512.	Each A100/H100 SM (streaming multiprocessor) is 4 way super-scalar and thus can execute 4 concurrent warps (32 threads). The threads in the warp are SIMT (Single Instruction Multiple Threads) can execute in a SIMD fashion or independently.
1.27	Could you please provide an assessment of the overheads introduced by the layered hierarchy of Grid, Clusters, and CTA? Additionally, how might these overheads affect compute efficiency at scale within the RAN context?	The overhead is increasingly higher as we move from CTA to Cluster to Grid scope due to the number of threads involved in the synchronization. Synchronization adversely affects compute efficiency since threads which complete work early are waiting for other threads to arrive. For the same reason it is always recommended to use synchronization at the lowest scope possible. In our experience with Aerial, the scope of synchronization is determined by the number of threads used to process a single user or a group of users. So scaling up of RAN workload i.e. number of cells is unlikely to increase the scope of synchronization.
1.28	Any updates on components, interfaces in cuPHY, cuMAC?	Please look into cuBB documentation and release notes.
1.29	Do you have the same capability and full features of the underlying CUDA libraries using just the Python API? What are the limits of using just the Python API?	Please look into CUDA Python support.
1.30	What cloud-native tools and strategies are being utilized to ensure scalability and fault tolerance in the Aerial Platform?How open is NVIDIA Aerial to third-party customization, and what collaborative models do you	Aerial can be installed and orchestrated via K8s. Please contact us for collaboration.

	envision for leveraging new AI-based 6G solutions?	
1.31	Could you please elaborate on how the size of batching might or might not increase latency and impact radio performance? What factors should we consider to achieve the right balance in the RAN context?	In general, batching is highly recommended for a throughput oriented platform like GPU. The factors which influence are the input dimensions, resource foot print of the implementation, available hardware resources, etc.
1.32	In the Graph Topology case, how should different PHYsical channels, Transmission of Periodic SSB, CSI-RS, PRACH etc. be handled?	If the physical layer channels (e.g. SSB, CSI-RS, PRACH, etc.) have different start and end times, we recommend creating and launching graphs per physical layer channel.
1.33	Will conditional nodes be supported for graphs launched from device?	Conditional nodes are not supported in device launched graphs. However, device-launched graphs can themselves be used to launch graphs conditionally.
1.34	L2 schedules the packet to L1 in GPU on download, L1 needs to buffer the packages, what's the max buffer size on the L1 side?	Presumably the question is about transport block buffer size which L2 provides L1 in the downlink. There are several factors which affect max buffer size: the max size per transport block, number of transport blocks which need to be supported, other L1 buffers in the system, etc. Additionally the GPU partitioning by MIG also matters. A Hopper GPU (without any MIG partitions) on GH200 contains 96GB HBM3 or 144GB HBM3e.
1.35	How is the use of GPU compared with multi-core or multi-thread CPU processing? If it's faster, why?	The GPU has a large number of small cores and works on SIMD architecture. It is good for parallel computing vs CPU is not.

1.36	How to debug issues such as TTI stretch, logging of channel level metrics?	One could use a combination of CUPTI and application level logging to debug the cause of TTI stretch
1.37	In the Aerial CUDA/GPU framework, how can we make sure processing received signals in one slot is complete in that slot?	You can use CUDAEvents to measure time spent in the GPU.
1.38	What is the best strategy to parallelize MIMO processing with multiple layers using the techniques mentioned - with best practices/visualizations, if any?	Please look into how cuBB implements Massive MIMO under release 24-3.
1.39	Any numbers on performance gains of using CUDA RAN compute, for different system configurations (UEs, Cell Params, Features), best and worst case latencies?	Please look into release notes of cuBB latest release which shows capacity can be handled by GPU. GPU focus is on DU L1 and L2 acceleration.
1.40	Are GPUs currently capable of handling RAN workloads? What are the limiting factors from the perspective of GPUs when it comes to managing RAN workloads?	Please look into release notes of cuBB latest release which shows capacity can be handled by GPU. GPU focus is on DU L1 and L2 acceleration.
1.41	Is GPUNetIO standardized within the industry ? Is it similar to DPDK ?	GPUNetIO is optimized to have NIC/GPU data transfer at lower latency, high throughput and low overhead on CPU.
1.42	What is the mechanism used to synchronize DU and RU and how is timing derived in Aerial?	DU is slave to timing from external fronthaul switches.
1.43	Any comments on TRP (Transmission Reception Point) numbers (8T8R, 32T32R)?	Aerial supports 4T4R. Support has also been added for 32T32R and 64T64R with different levels of test coverage. Please contact us if you have further questions.

1.44	Do you support the O-RAN 7-2 interface or CPRI (Option 8) interface?	ORAN 7-2. It is software defined architecture that can support any pipeline.
1.45	Are these based on Dell/Gigabyte server or Grace Hopper server?	It can be exercised both on the X86 and Grace Hopper system. For a performant system, Grace Hopper is recommended.
1.46	Have you exercised AI/ML approaches into cuPHY or cuMAC?	Yes, we plan to release that work in the year 2025.
1.47	Does cuPHY support Massive MIMO 64T64R?	Yes, with release 24-3.

Section 2: Aerial Omniverse Digital Twin

Aerial Omniverse Digital Twin		
	Questions	**Answers**
2.1	In a sentence or two, can you define what a RAN Digital Twin is? How is it different from the automation of a traditional network design tool that uses either ray tracing or statistical models to generate RF coverage maps?	The digital twin is a single framework to cover network development, deployment and operations. Due to this, its requirements go beyond those of a system-level simulator, or a network planning tool.
2.2	What are the challenges and best practices identified by NVIDIA for scaling digital twins to simulate large RAN deployments (e.g., 57 cells with 5700 UEs)? How does the Omniverse platform ensure accurate RF propagation modeling at this scale?	Multi-GPU is necessary to reach these scales; also critical bottleneck areas need to be carefully coded in CUDA thinking about scale and performance. Our RF model is differentiable, thus it can be tuned against measurements. We hope that the tuning is capable of removing major inconsistencies against measurements. More work is necessary to explore the boundaries of what can be tuned out and is intrinsically a twinning issue.
2.3	How easy is it if we want to adapt it to some research that requires changing some low-level modules, e.g., for new MIMO algorithms, for	At the moment (release 1.1), the coding would have to happen in C++. In the first half of 2025, this will be possible in accelerated Python.

	simulating metamaterial surfaces, etc.?	
2.4	Can this be used for all 6G frequency bands? At mmWave, isn't it so that only direct Line Of Sight connections are relevant?	The EM model has no issues above 100MHz. At different bands, the requirements on the accuracy of the map will be different though. That's why we are introducing a differentiable model, so that, in the presence of measurements, the EM model can be tuned and the discrepancies against the field can be minimized.
2.5	Does the ML based channel estimation use pilot symbols/ reference signals?	Yes, fully interchangeable with the MMSE counterpart.
2.6	What is the underlying framework? Is it Sionna?	For AI, it's Torch, for PHY and MAC is NVIDIA cuPHY and cuMAC. UI is based on NVIDIA Omniverse KIT. The rest is AODT code.
2.7	Are there any unique features in Sionna that made it a particularly good fit for this project?	Sionna allows to create designs for the data plane by providing a link level environment. Evaluating the performance of a design obtained with link-level simulations in a system-level environment (in the same way in which RAN1 intends "system-level environment") is a natural progression in the maturation of data-plane designs for L1. We consider Sionna and AODT to be well aligned and synergetic in nature.
2.8	Is Sionna completely open source? Meaning we can see the algorithms implemented for the different features of the phy?	AODT is free to use for non-commercial purposes. For commercial purposes, it requires a license.
2.9	How fast can the simulation run - faster than real time?	Current runtime is 500ms per slot per cell on L40S. Target is roughly 1000x faster on a single L40S.
2.10	What are the hardware requirements for running Aerial Omniverse Digital Twin?	Details on the installation process for the Digital Twin can be found here: https://docs.nvidia.com/aerial/aerial-dt/text/installation.html#id1
2.11	What are the details of Physical Layers channel configurations supported?	Currently, anything supported by cuPHY. Please refer to:

	Multi Cell scenario with different Mobility.	https://docs.nvidia.com/aerial/cuda-accelerated-ran/aerial_cubb/aerial_cuphy/features_overview.html for more details.
2.12	What is the maximum number of gNB and UE nodes that can be emulated at a given moment?	Our target is to have 9 cells and 900 RRC connected UEs in a single L40S GPU, with the final constraint that the simulation should not exceed 10x real time. Currently, it's possible to exceed this number even with 1 GPU, but the performance needs to be improved.
2.13	Propagation models seem to be purely theoretical. What is the plan to finetune the prop models?	In the first half of 2025, we are going to have a differentiable version of the model, which can be tuned against measurements using backpropagation.
2.14	How does the RAN controller coordinate data flow and decision-making between physical and logical layers in simulated networks?	Such decisions are not taken by the controller, but by L2 and L3 on the RAN side. The controller simply orchestrates L1, L2 and L3.
2.15	Please explain how the simulation clock functions in RAN simulation.	As per release 1.1, everything is slot-aligned. From early 2025, it'll be possible to have temporal misalignment between serving and interfering links. In any case, temporal granularity is in the FFT sample.
2.16	Since this is not real time simulation, how do you ensure random access procedure is happening like exact real world deployment? Furthermore, can I see scenarios like random access transmission failures (UE can hear gNB but not the other way around)?	The simulator is currently event-based. So, even if it's not real time, no timer will run out and the full logic can be captured.
2.17	Do you have considerations for different types and densities of vegetation?	Our first approach with vegetation will be procedural. With that, we intend to make the densities, heights, and other qualities of the vegetation configurable. There is of course a lot that can be done there, and we will be exploring it/remain open to feedback.

2.18	With features like vegetation already available in version 1.4, what criteria guide the inclusion of elements like the troposphere and ionosphere, especially for expanding coverage in challenging environmental conditions?	Troposphere and ionosphere effects are for non-terrestrial networks.
2.19	Are you building ORAN or a simulation too? I didn't understand the objective. This is one of many types of implementation. So how does this become a digital twin?	The design phase is only the first aspect that the tool needs to support. Network planning and feature testing before roll out is the next milestone, and finally L3 needs to be correctly parametrized and tuned based on location. The joint support of all three aspects will make this a digital twin. Clearly, the effort to get there is substantial.
2.20	With so many equalizer algorithms available, what drives the choice between RZF, noise-diagonal MMSE, and MMSE IRC for specific network environments?	It's the balance between link performance and runtime performance. Depending on the makeup of the interference and the necessity to hit a specific BLER target, we can adjust the underlying algorithm as we see fit. This kind of consideration comes to mind when the latency of the algorithm is relevant. Otherwise, state of the art is the solution to obtain pareto-optimal solutions.
2.21	Do EM models support subTHZ, THZ, mmWave bands?	The EM model is asymptotic, meaning that it becomes increasingly accurate as we go to higher and higher frequencies. The difficulties with higher frequency bands is the level of detail with which the map needs to be portrayed. We consider "Twinning for THz" an open problem. That is, more polarimetric field measurements are needed to extend operations in such bands and retain claims of realism.
2.22	Any hints on how radio calculations are made while being faster than real time? Based on TR 38.901 or otherwise?	We are using geometrical optics and extensions in conjunction with NVIDIA HW raytracing capabilities.
2.23	How is the covariance of the interference computed?	Different algorithms would compute things differently. Papers on the use of shrinkage to

		improve the estimate would provide some background here.
2.24	What types of RAN are supported in this simulator? Is it focused on 3 sector macro, or can more complex (INBLD, IDAS, many sector, etc.) configurations be simulated?	We are targeting the possibility of having 100s of sectors, with tens of thousands of UEs. Performance targets are defined for 57 cells x 5700 on 8x L40S GPU and 9 cell x 900 UEs for 1x L40S GPU.
2.25	Why is the scene unit in cm and not in meter?	The Omniverse Kit, on which we've built the UI, defaults to centimeters.
2.26	Are there different levels of speed for mobility of UE?	UE speed can be changed freely; currently there is no way to assign a speed category, but it's a good idea and we can look into that.
2.27	Can velocity vectors vary based on surrounding network density?	We plan to allow any type of mobility in any density. This kind of change, if of interest, can be easily added to the code we make available.
2.28	Do the UEs and nodes have fixed height? Does the height use flexible parameters?	RU height can be changed individually and UE heights can be changed as a group (meaning that all UEs have the same height). This will be addressed in future releases.
2.29	Can you provide simulation time needed for a typical multi-cell multi-UE configuration?	Current runtime is 500ms per slot per cell on L40S. Target is roughly 1000x faster on a single L40S.
2.30	Can we use lidar/cloud points to generate gml/scene?	Yes. Right now there is no official support, but it could be done using 3rd party tools. In future releases we hope to have more clearly defined workflows for this.
2.31	How do you model the remote electrical tilts?	This would be through a static beamforming codebook added to the panel. We model every antenna element individually, so it is not an issue to add static networks of phase shifters.
2.32	Can non UEs move around in the mobility domain (like Trucks or other passive objects)?	At the moment it is just UEs that can move around the mobility domain. Having other moving objects (trains, vehicles) is coming in the first half of 2025.

2.33	Why does UE need a transmit beam former since it has only 1 /2 antennas?	UE can have a much larger number of antennas. Use case is not limited to phones.
2.34	Is Digital twin capable of supporting future use cases like ISAC?	ISAC is important. Within the first half of 2025, we are going to add dynamic scattering and back scattering and ideally some initial version of ISAC based on OFDM.
2.35	Does digital twin support 5G NR NTN use cases? If yes, any demos?	There is no NTN support for the moment. This will be prioritized based on request.
2.36	With NTN and satellite companies deploying 5G based networks, do you see them being able to use this digital twin framework? Does it have features built in for this use case?	This would require adding tropospheric and ionospheric effects; we are aware of the work that needs to be done. We will prioritize based on the interest of the developer community.
2.37	Do you support Aerial drone models?	Supporting photogrammetry maps is very interesting and we hope to create some tools for this in the future. We've had some successes working with this type of data in Blender and then converting to USD (plus adding template properties to the map). That is the workflow we'd recommend right now, though it should get easier with future releases!
2.38	In your opinion, can we get a scene between two cities using CityGML or is it limited to a single city setup?	It depends how close the cities are. If they are close, you may have a scene with two or more cities. We recommend scenes under 200km x 200km.
2.39	Is it possible to generate a Radio Coverage Map using CityGML?	There is no technical problem in doing so, but the functionality to produce coverage maps will be enabled in a second stage.
2.40	Is Blender supported for scene creation from OSM or only CityGML?	We offer two tools you may use. 1) OSM import, which pulls OSM directly. With this we only need to know the bounding box of interest. 2) CityGML import. If you have other data you'd like to bring in the scene and we don't directly support it, you can do so yourself with most 3rd party tools that can

		compose USD. For this latter point, we will be releasing more tools/examples to help.
2.41	In an earlier discussion of scene creation, the scene is created just by taking area? And did you use any open source data like building or osm data for roads?	For our OSM import tool, the scene is created from a bounding area. We pull OSM data for that workflow.
2.42	Can you use Cesium ion 3D Tile structures?	Not at this time, but we are very interested to add support for 3dTiles.
2.43	Is it possible to provide our own channel model per location to be used in the simulations of both EM and RAN. The channel model is based on real measurement, probably through the channel emulator block.	Yes, it is possible. We offer a clear and documented interface to the EM propagation model. It can be replaced with any other model.
2.44	Could we develop materials that 'adapt' their EM properties on demand?	This kind of thought will be explored in the scope of RIS, but it's not yet in sight for us at the moment. Happy to learn more about the use case to see if we can prioritize.
2.45	With EM signals often reflecting, scattering, or diffracting in dense cityscapes, could fractal models, which capture natural randomness, better anticipate signal distortions in complex areas?	We expect this kind of idea to become a bit more concrete where full-wave simulations become necessary.
2.46	Are there any considerations on security mechanisms and protocols in RAN? How about testing multiple attacks on the DT platform and training AI models for anomaly detection?	That requires more of L3 in place. We have taken note of it and it will come depending on priorities.

2.47	Can you simulate a RIC making real time changes, say to frequency band?	That requires more of L3 in place. We have taken note of it and it will come depending on priorities.
2.48	Can performance analysis across multiple architectures be evaluated within the Digital Twin? I.E. - KPI monitoring with PNF vs VNF architecture to decide where to deploy a limited amount of virtualized assets?	That requires more of L3 in place. We have taken note of it and it will come depending on priorities.
2.49	Can we have the RRC logs when running the AODT simulation?	L3 is only sketched out for the moment, we don't have all of L3 in place.
2.50	Could quantum tunneling allow signals to pass through materials once considered impenetrable?	Such events would be too rare to affect a sampled environment.
2.51	The configuration mentions MMSE and Neural Network options for channel estimation in the PUSCH pipeline. What performance differences have you observed between these methods?	This will be explained in an upcoming publication.
2.52	Is there a comparison between the propagation analysis results from DT simulation and the real world?	We are working towards it.
2.53	Is there any validation conducted between the raytracing result and real life measurement?	We are working towards it.
2.54	Can you compare your predictions to actual measurements?	We are working towards it.

2.55	Can we add equivalent sources such as surface sources like Huygens sources?	Net yet possible, but can be added if interest is high enough.
2.56	How do you configure multiple CP lengths based on symbols?	Not yet possible at the moment; will be added in a future release.
2.57	Was any sensitivity analysis done to compare ray-traced channel estimates with ground truth from real world captures and BLER differences between the two?	We are working towards it.
2.58	Is CA (carrier aggregation)/ NW slicing supported?	Not at the moment; implementation will come depending on interest. Noted.
2.59	Is there no downlink csi from the UE?	It's scheduled for 1.3.
2.60	How can I work with a map that is already a mesh, such as an OBJ file?	That is possible with a bit of work. We recommend using a tool like Blender to prepare your scene. You will need to do a few things, like separate your ground and building meshes, ensure there are no non-manifold geometries, and triangulate all faces. You will also need to add the required primary attributes to each layer (ground plane, buildings, mobility mesh). Then export to USD. We plan to offer more tools to support this workflow in upcoming releases.
2.61	Will 3D Tiles be supported?	Yes, we hope to add 3D Tiles support in future releases.
2.62	What is the maximum map size that is supported?	At this time, we recommend maps that are maximum 200kmx200km.
2.63	If the users are interested, is it possible to use a different EM propagation model with the Aerial Omniverse Digital Twin?	Yes, this is possible. AODT documentation (specifically here: Additional Information - NVIDIA Docs) indicates the function calls that the internal EM solver executes. If any other EM solver is

		wrapped by a class that respects the API, there is no problem in integrating a different EM solver.
2.64	What are NVIDIA's plans to support statistical channel models?	38.901 CDL coming in the first half of 2025.
2.65	From the description, a good part of the logic is in the controller. In the scope of 1.2, will the introduction of Python be restricted to the PHY, or the controllers will also be provided in Python?	The first step towards Python will be the main controller. We are currently still testing whether we want to have this shipping with 1.2.
2.66	How far away from real-time is the simulation? Is it already in the range of "not slower than 10x real-time"? Or is it slower now?	By design, runtime scales linearly with the number of RUs, and very sublinearly with the number of UEs. Single RU is between 100x and 1000x away from real time on a single L40. Multi-GPU setup (coming in the second half of 2025) will help, but we need more performance optimization to get to 10x from real time.
2.67	Is multi-GPU operation already available? When will it become available?	Second half of 2025.
2.68	As energy efficiency becomes a major factor in scaling AI and network workloads, how does the Omniverse platform incorporate features for dynamic resource partitioning to optimize GPU utilization across different types of workloads?	Resource partitioning won't be necessary until real time is reached. First we will need to bring the platform to realtime. After that's achieved, green contexts are likely going to be the main method for SW-level resource partitioning.
2.69	How can we ensure consistency between the physical and digital twins, considering frequent RADIO changes (topology, configuration, status, etc.)?	Tuning effort of the EM solver is ongoing: field measurements (channel sounder, phone, spectrum analyzer etc.) can be used to tune the EM solver using backpropagation.

2.70	Any support to simulate very high speed scenarios like high speed trains?	The EM model is already capable of that. By manually defining the mobility of a fleet of UEs, this scenario is already possible in AODT. More utilities to make things easier will come in the course of 2025 though.
2.71	Can a custom antenna radiation pattern be defined through radiation pattern measurement?	Yes, this is possible.
2.72	Did you use the 3GPP channel model for path loss calculation?	No, propagation is not based on a channel model. Propagation occurs as per a simplified version of Helmotz's equation. This includes deterministically characterizing scattering events from objects.
2.73	How were the models validated and verified?	Ongoing effort: Field measurements (channel sounder, phone, spectrum analyzer, etc.) can be used to tune the EM solver using backpropagation.
2.74	For future enhancement of the scene, are you considering any street furniture like street poles, stop signs and building facade?	We will consider this in the scope of tuning against field measurements.
2.75	Is there support for 5G UL SRS?	Yes.
2.76	Can we have our own PHY solution running on DT rather than Nvidia's PHY solution?	Yes, as long as it is SW-based.
2.77	What factors drive the choice of maximum cells per slot and per block parameters in your RAN simulation?	vRAM size is ultimately the limiting factor.
2.78	How can we reach out for sharing thoughts and features?	We recommend joining the 6G Developer Program, which then allows you to post in the Aerial forums.

Section 3: Aerial AI Radio Frameworks

Aerial AI Radio Frameworks		
	Questions	**Answers**
3.1	What type of GPU do I need to use Sionna?	Sionna runs on essentially any GPU and CPU.
3.2	Can I use Sionna in commercial applications?	Sionna is distributed under the Apache 2.0 License which allows for commercial use.
3.3	How can I contribute to Sionna? What's the best procedure for external contributions?	File pull requests on GitHub.
3.4	In wireless communications, not all algorithms are differentiable. How does Sionna handle non-differentiable algorithms?	Custom gradients can be provided for any processing block. Alternatively, there are gradient-free optimization algorithms that can be used for training.
3.5	Can I use Sionna for indoor scenarios?	Yes
3.6	What are some of the challenges for bringing machine learning to wireless?	In the context of ML in the physical layer, one consideration is the latency of
3.7	Does Sionna provide an ability to model Integrated Access and Backhaul?	Yes, but not out-of-the-box.

3.8	Our research group is interested in equivalent source-based ray tracing with Huygens principle. Does Sionna allow low level ray trace mechanism control (like launching-secondary rays from Huygens sources?)	In principle, yes, but it would require significant development effort.
3.9	I am interested (for the higher layers) in generating SINR matrices between many (possibly hundreds) of users. Is Sionna a suitable tool for that?	Yes. However, for very large systems you need a GPU with sufficient memory.
3.10	Does Sionna support Coordinated Multipoint?	Yes. Antennas can be arbitrarily distributed.
3.11	Does Sionna have support for integrating the antenna patterns that do not have closed from E field expression but the E field measurements from an EM simulator?	Yes, you can provide a custom antenna pattern.
3.12	Is Sionna completely open source? Meaning we can see the algorithms implemented for the different features of the phy?	Yes, Sionna is open source.
3.13	When modeling UE mobility with Sionna ray tracing based channel modeling, how can we define valid user trajectories?	There is no built-in mobility model that creates valid trajectories.
3.14	Is it possible to simulate a multi-cell multi-user MIMO communication system using the TR 38.901 stochastic channel models in Sionna? How can we	Yes. You can simulate interference by simply adding additional channels modeling interfering transmissions.

	model the uplink intercell interference in such a scenario? Thank you very much in advance!	
3.15	Any opportunities for Sionna, Aerial, and RAN Digital Twin to seamlessly integrate? Somehow they seem to implement three separate workflows and the integration is an after thought. Thanks.	For now, there is no seamless integration between the tools.
3.16	Does pyAerial use torch/Sionna(TensorFlow) for AI components and cuPHY for the conventional components?	pyAerial currently makes use of cuPHY only; in release 1.2, it will support JIT python and we will add Torch support for AI.
3.17	So can we say that pyAerial is the commercial version of Sionna that can utilize CUDA ?	Both pyAerial and Sionna use CUDA. Sionna is made for rapid prototyping while pyAerial is optimized for performance.
3.18	Is it required to have cuPHY code to use pyAerial model for experimentation? Or can pyAerial run on CPU as well?	pyAerial is a library that provides a Python API to the Aerial cuPHY kernels. Each time a pyAerial component is executed, for example a call to the pyAerial LDPC decoder, the CUDA code for the component is executed. Therefore a GPU is required to run a pyAerial model. The cuPHY code is also required.
3.19	Is pyAerial suitable for real-time signal processing?	pyAerial is sitting on Python. It provides a Python API to CUDA wireless signal processing kernels from which the Aerial CUDA accelerated RAN is built. So while the CUDA kernels are heavily optimized for real-time performance, the overhead of Python means that you cannot really run a pyAerial model in real-time. So at this point you can think of a pyAerial model as being a simulation.

Section 4: ARC-OTA

ARC-OTA		
	Questions	**Answers**
4.1	When you say 6G testbed, which 3GPP release does it support?	Good Question, there is no 6G standard yet. We are positioning this for research organizations working towards defining 6G.
4.2	Do you support O-RAN 7-2B today?	Yes, we support O-RAN 7.2B split for massive MIMO.
4.3	Are you able to run aerial DT in the ARC-OTA 1.5?	ARC-OTA testbed can run Aerial CUDA Accelerated RAN and Aerial AI Frameworks. Aerial Omniverse Digital Twin will need RTX-based GPU, which is not the type we used in ARC-OTA.
4.4	Do you have FR1 vs FR2 feature differentiation in your offering?	Our current offering supports FR1.
4.5	How does ARC-1 (introduced here: https://developer.nvidia.com/blog/bringing-ai-ran-to-a-telco-near-you/) relate to ARC-OTA and the others you described?	ARC-1 was recently announced during TMO Capital Day by our CEO, Jensen. ARC-1 stands for Aerial RAN Computer-1, which is a core component that enables the AI Aerial platform i.e., converged infrastructure that allow telcos to do both GenAI & SW-defined RAN workload. ARC-1 is

		based on GB200-NVL2 (i.e., 2 GB200 superchip connected via NVL) and BF3 DPU, preloaded with CUDA and DOCA OS. ARC-OTA is a NVIDIA 6G Research Testbed. It is an E2E OTA network in the loop that consists of gNB server, GM, FH-switch, O-RAN compliant O-RU, 3GPP compliant 5G commercial UE or OAI based soft UE. It is preloaded with NVIDIA Aerial CUDA-Accelerated RAN and OAI L2+ & OAI 5GC to offer world 1st full-stack SW-programmable wireless research testbed.
4.6	Our primary interest is in FR3 and FR1 with massive MIMO use cases. Would be interested in learning more about your solution	Great! Please join NVIDIA 6G Developer Program to get access to all the assets: https://developer.nvidia.com/6g-program
4.7	Do you have detailed instructions for setting up the ARC-OTA software stack? Also, is the full stack available for download?	Yes, the detailed instruction for ARC-OTA setup is available here: https://docs.nvidia.com/aerial/aerial-ran-colab-ota/current/index.html. The full SW with source code can be downloaded once you join NVIDIA 6G Developer Program.
4.8	Which model is used in LLM?	Llama 70B with Retrieval Augmented Generation (RAG) using NVIDIA's TensorRT-LLM and RAPIDS libraries.
4.9	How many UEs can be used in a test bed?	We have tested 8 CUEs in an OTA environment.
4.1	Can we use our Gen 1 development kit to build the latest ARC-OTA environment?	Gen 1 is supported but will be End of Life in 2025.
4.11	How much does it cost to deploy an on-campus network for academic institutions?	It really depends on the capabilities of the on-campus network such as Use case, # of Cells, etc. The items to consider are: GH200+2xBF3 DPUs,xHaul switch, GPA antenna, GPS antenna installation, Grand Master, O-RU, devices, equipment rack. Then you

		need to consider the SFPs and optical cables to connect the network equipment. This type of question is best answered if there is a specific deployment use case being considered.
4.12	If I'm interested in ARC-OTA, how can we get started?	All information on how to get started with ARC-OTA can be found here: https://docs.nvidia.com/aerial/aerial-ran-colab-ota/current/index.html
4.13	Where can I get more info on the OAI stack?	If you click on Getting Started in the ARC-OTA documentation page, it has an OAI link and how to get started from HW & SW BOM.

Appendix A. NVIDIA AI Aerial FAQ

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

Arm

Arm, AMBA, and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore, and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS, and Arm Sweden AB.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Copyright

© 2024 NVIDIA Corporation. All rights reserved.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com