**III. AI Toolkits and Potential Applications for Mobile Networks**

1. **AI Toolkits**

NVIDIA provides several advanced AI toolkits that accelerate the design, testing, and operation of current and future wireless communication systems, including 5G and 6G. These include the NVIDIA Aerial SDK, the NVIDIA Aerial Omniverse Digital Twin (AODT), and Sionna.

Here are the definitions of these AI toolkits:

* **NVIDIA Aerial SDK:**  The NVIDIA Aerial SDK is a **collection of software-defined libraries** specifically optimised to execute **5G gNB workloads on the GPU**. It is designed to accelerate 5G and 6G Radio Access Networks (RANs) on NVIDIA's GPU and DPU hardware, built to be **cloud-native** and integrate seamlessly with AI/ML frameworks. It simplifies the creation of highly programmable and scalable software-defined 5G RANs using commercial off-the-shelf (COTS) servers with NVIDIA GPUs.

The core components of the Aerial SDK are:

* + **CUDA Baseband (cuBB):** Regarded as the **"heart of Aerial"**, cuBB provides a **GPU-accelerated 5G signal processing pipeline**. It incorporates **cuPHY for Layer 1 (Physical Layer)** and **cuMAC for the Layer 2 (Medium Access Control) scheduler**, ensuring high throughput and efficiency by keeping all processing within the high-performance GPU memory. cuBB is software-defined, scalable, modular, highly programmable, and cloud-native, intentionally avoiding fixed-function accelerators.
    - Within cuBB, **cuPHY-CP (Control Plane)** acts as the control-plane software that bridges Layer 1 cuPHY with upper layer stacks (like L2/L3) and the O-RAN fronthaul interface. It orchestrates L1 functionalities, handles slot information, and manages GPU resources. The **cuPHYController** is the central component that manages the entire 5G Layer 1 (PHY) processing, orchestrating data flow between the network and the GPU.
  + **CUDA Virtual Network Functions (cuVNF):** This SDK works in conjunction with cuBB by providing **optimised input/output (IO) directly to GPU memory** from GPUDirect-capable Network Interface Cards (NICs), such as Mellanox CX-5 and CX-6.

The overall **Aerial CUDA-Accelerated RAN** framework is built on a foundation of GPUs, DPUs, and CPUs, operating within a cloud platform that integrates an Operating System, Kubernetes (K8s), and Operators. It leverages NVIDIA's CUDA programming model for GPU acceleration and DOCA for DPU programming. This framework is designed for interoperability with third-party Centralised Units (CU) and RAN Intelligent Controllers (RIC).

* **NVIDIA Aerial Omniverse Digital Twin (AODT)** The NVIDIA Aerial Omniverse Digital Twin (AODT) is a **next-generation, system-level simulation platform** used for cutting-edge research and development on 5G and 6G wireless systems, as well as for optimising network planning and operations. AODT applies **ray-traced channels** to the physical (PHY) and medium access control (MAC) layers of the NVIDIA Aerial CUDA-accelerated RAN platform. It simulates the system-level performance of an actual network deployment without abstractions. Channels are computed using high-performance ray tracing of detailed models of the real world, and it runs the software-defined NVIDIA Aerial CUDA-accelerated RAN platform.

AODT is a **unique tool to benchmark system performance and explore machine learning-based wireless communication algorithms** under real-world conditions. It enables the extraction of data at any point in the communication protocol stack for training RAN data that would otherwise not be available, making it instrumental for integrating ML and AI into 6G networks. The platform is **built on NVIDIA Omniverse**, which facilitates a convenient graphical user interface, visualisations, and remote collaboration for distributed teams.

* **Sionna** Sionna is an **open-source, GPU-accelerated, and differentiable library** designed for research in communication systems. It features a fast ray tracer for radio propagation, a versatile link-level simulator, and capabilities for system-level simulation. Sionna is considered an **essential tool for advancing future communication systems, such as 6G**.

Sionna is built on top of powerful **automatic differentiation frameworks**, specifically **TensorFlow** for its Sionna PHY and Sionna SYS modules, and **Mitsuba 3 and Dr.Jit** for Sionna RT. This foundation enables gradient-based optimisation and machine learning integration. It provides a **high-level Python API** that simplifies the modelling of complex communication systems while ensuring modularity and extensibility, with every building block being an independent module.

Sionna includes three core modules:

* + **Sionna RT:** A standalone ray tracer for modelling radio propagation. It supports detailed scene loading, visualisation, ray tracing, converting paths to channel responses, and includes features like Radio Maps and mobility.
  + **Sionna PHY:** A link-level simulator for wireless and optical communication systems. It offers a wide range of functionalities including FEC (LDPC, Polar), Mapping (Constellation, Mapper), various Wireless Channel Models (AWGN, 3GPP 38.901), OFDM tools, and MIMO utilities.
  + **Sionna SYS:** Provides system-level simulation using physical-layer abstraction. It includes modules for PHY Abstraction, Link Adaptation, and schedulers like Proportional Fairness Scheduler, along with tools for multicell topology generation (e.g., Hexagon, HexGrid) [i, 40, 45, 54]. Sionna SYS can also be integrated with Sionna RT for advanced, ray-tracing informed system-level simulations.

A significant new component introduced in Sionna 1.1.0 is the **Sionna Research Kit (SRK)**. The SRK enables the **deployment of trained AI/ML components into a real software-defined 5G NR radio access network**. It is built upon the OpenAirInterface project and is powered by the NVIDIA Jetson AGX Orin platform.

Here are the minimum specifications for NVIDIA Aerial Omniverse Digital Twin (AODT), NVIDIA Aerial SDK, and NVIDIA Sionna.

**NVIDIA Aerial Omniverse Digital Twin (AODT)**

There are specific GPU and vRAM requirements depending on the deployment type, along with recommended operating systems and examples of qualified systems for CPU and memory.

**GPU and vRAM Requirements**:

* **Frontend alone:**
  + **1 NVIDIA GPU**
  + **12GB+ GPU vRAM**
  + GPU Requirement: GTX/RTX (e.g., RTX 6000 Ada, A10, L40)
* **Backend alone:**
  + **1 NVIDIA GPU**
  + **48GB+ GPU vRAM**
  + GPU Requirement: (e.g., RTX 6000 Ada, A100, H100, L40)
* **Frontend and backend replay (single GPU setup):**
  + **1 NVIDIA GPU**
  + **48GB+ GPU vRAM**
  + GPU Requirement: (e.g., RTX 6000 Ada, L40)
* **Frontend and backend colocated (concurrent operation):**
  + **2 NVIDIA GPUs**
  + Requires **1x frontend-capable GPU and 1x backend GPU** (vRAM requirements as above)

**GPU Driver Versions** (used by installation scripts):

* Frontend (Linux): 550.127.05
* Frontend (Windows): 552.55
* Backend (Linux): 560.35.05
* Frontend and backend replay (Linux): 560.35.05
* Frontend and backend colocated (Linux): 560.35.03

**Operating System Support**:

* Frontend alone: **Windows 11, Windows Server 2022, Ubuntu 22.04**
* Backend alone: **Ubuntu 22.04**
* Frontend and backend replay: **Ubuntu 22.04**
* Frontend and backend colocated: **Ubuntu 22.04**

**Supported Streaming Multiprocessor (SM) Architectures** for the AODT backend: **80, 86, 89, and 90**. The backend performs startup verification to ensure it runs on a supported SM architecture and checks for compile-time vs. run-time architecture alignment.

**CPU, Memory, and Storage:** It is recommended to look at **qualified systems** for these requirements. Examples of qualified systems include:

* **Azure VM (Multi-Node):**
  + Frontend Node: **36 vCPUs, 440GB Memory**
  + Backend Node: **24 vCPUs, 220GB Memory**
* **Dell R750 (Colocated):**
  + **Intel Xeon Gold 6336Y 2.4G, 24C/48T CPU**
  + **512GB DDR4 Memory**
  + **2TB Storage**

AODT is designed to simulate from small-scale deployments on a **single GPU** to city-scale models on GPU clusters. It can be accessed on premises, on laptops, via public cloud, or through NVIDIA cloud service.

**NVIDIA Aerial SDK**

The Aerial SDK runs on **NVIDIA-certified EGX servers** and requires a specific software stack. The website for NVIDIA Aerial SDK provides "Software Manifests" and "Overall Platform Qualification" details, which outline the supported and validated configurations rather than strict minimums.

**Software Requirements**:

* **Host OS:** **Ubuntu 20.04 with 5.4.0-65-lowlatency kernel** (for Aerial SDK workload) or **Ubuntu 20.04 server** (for Kubernetes stack).
* **Container OS:** **Ubuntu 20.04**.
* **CUDA:** **CUDA 11.7 Toolkit**.
* **Host GPU Driver:** **515.43.04+** (for Aerial SDK workload) or **R515** (for Kubernetes).
* **NVIDIA-peermem:** Same as CUDA driver version.
* **Mellanox OFED:** **5.7-1.0.2.0**.
* **CX6-DX NIC Firmware:** **22.34.1002**.
* **A100x Firmware:** **24.34.1002**.
* **DPDK:** **DOCA DPDK**.
* **Containerd:** **1.5.8** (for Kubernetes).
* **Kubernetes Version:** **1.23** (for Kubernetes).
* **Helm:** **3.8** (for Kubernetes).
* **NVIDIA GPU Operator:** **1.11.1+** (for Kubernetes).
* **Optional Software:** Matlab.

**Hardware Considerations** (based on qualified platforms):

* **GPU:** Optimised to run 5G gNB workloads on **NVIDIA GPUs**. **A100x and A100 GPUs** are explicitly qualified and support up to 4 carriers. Multi-GPU platform support (2 cells on each GPU) is also noted.
* **DPU/NIC:** Compatible with **GPUDirect-capable Network Interface Cards (NICs), such as Mellanox CX-5 and CX-6**. Mellanox CX6-DX is consistently mentioned in qualified platforms.
* **CPU:** Qualified platforms include **Intel Xeon Gold 6240R, 2.4GHz, 24C48T** and **Intel Xeon Gold 6336Y 2.4G, 24C/48T**.
* **Memory (RAM):** Qualified platforms specify **96GB DDR4** (Gigabyte E251-U70) or **512GB DDR4** (Dell R750).
* **Storage:** Qualified platforms list **480GB LiteOn SSD** or **2TB storage**.
* **PCIe Topology Requirements** are also mentioned as important.

**NVIDIA Sionna**

Sionna is designed for communication systems research and offers flexibility in hardware execution.

**Hardware Requirements:**

* **GPU:** Sionna is **GPU-accelerated** and **supports NVIDIA GPUs** for significantly faster simulations.
* **CPU:** Sionna **can also run on a CPU**. If running Sionna RT on a CPU, **LLVM is required** by Dr.Jit.
* **Sionna Research Kit (SRK):** This specific component of Sionna is designed to run on the **NVIDIA Jetson AGX Orin platform**.

**Software Requirements**:

* **Python:** **3.8-3.12**.
* **TensorFlow:** **2.14-2.19** (required for Sionna PHY and Sionna SYS modules).
* **Operating System:** **Ubuntu 24.04 is recommended**.
* **Environment:** Using a **Docker container or a Python virtual environment is highly recommended**.
* **Sionna RT:** Has the same requirements as Mitsuba 3, and refers to its installation guide for further information.

References:

* + <https://developer.nvidia.com/aerial-omniverse-digital-twin>
  + <https://www.youtube.com/watch?v=J5-rkgL2dFA>
  + <https://docs.nvidia.com/aerial/aerial-dt/text/installation.html>
  + <https://docs.nvidia.com/aerial/aerial-dt/text/overview.html>
  + <https://docs.nvidia.com/aerial/archive/aerial-sdk/22-4/text/product_brief/aerial_sdk_platform.html>
  + <https://developer.nvidia.com/sionna>
  + <https://nvlabs.github.io/sionna/installation.html>

1. **Technical Feasibility**