Aerial cuPHY Developer Guide

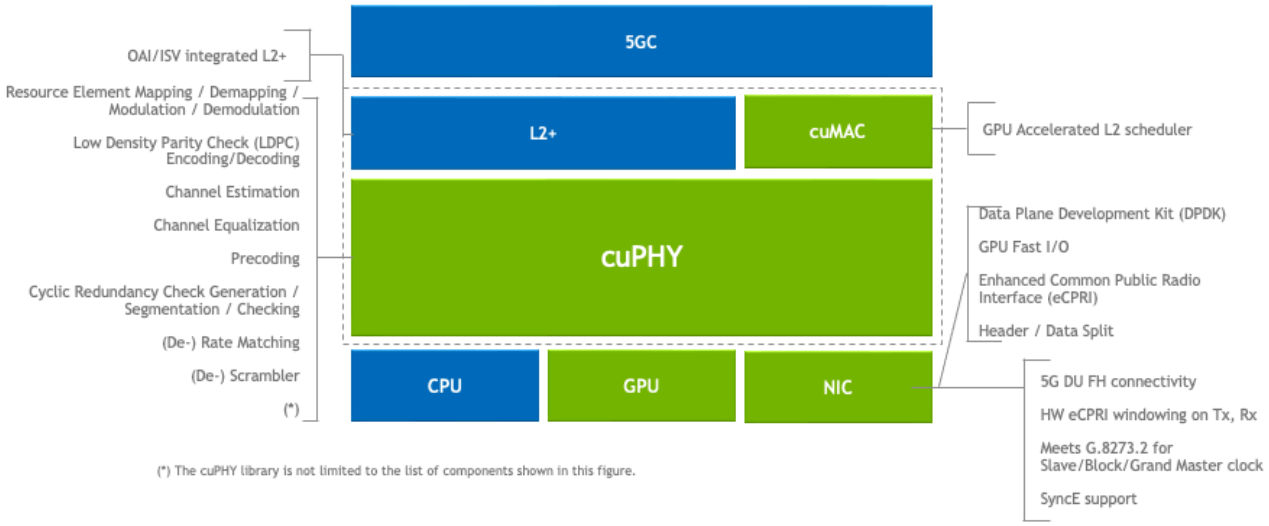
Table of contents

Aerial CUDA-Accelerated RAN is a set of software defined libraries that are optimized to run 5G gNB workloads on GPU. These libraries include cuPHY, cuMAC and pyAerial. In this section, we focus on layer-1 (L1), or physical (PHY) layer of 5G gNB software stack as defined by 3GPP [1-5].

cuPHY is the 5G L1 library of the Aerial CUDA-Accelerated RAN. It is designed as an inline accelerator to run on NVIDIA GPUs and it does not require any additional hardware accelerator. It is implemented according to the O-RAN 7.2 split option [8]. cuPHY library takes advantage of massively parallel GPU architecture to accelerate computationally heavy signal processing tasks. It also makes use of fast GPU I/O interface between the NVIDIA Bluefield-3 (BF3) NIC and GPU (GPU Direct RDMA [7]) to improve the latency.

BF3 NIC provides the fronthaul (FH) connectivity in addition to the IEEE 1588 compliant timing synchronization. The BF3 NIC also has a built-in SyncE and eCPRI windowing functionality, which meets G.8273.2 timing requirements.

In the following, we first give an overview of cuPHY library software stack. cuPHY library consists of L1 controller components running on the CPU and PHY layer functions running on the GPU. After providing the overview, we will go into details of each component and explain how L1 controller components interact with each other and L2. Finally, we will go over the PHY layer signal processing functions, which are accelerated as CUDA kernel implementations.



Aerial CUDA-Accelerated Software Stack within 5G gNB DU

Acronyms and Definitions

Acronym	Description
3GPP	Third Generation Partnership Project
5G NR	Fifth generation new radio
CB	Code Block
CSI	Channel State Information
CSI-RS	Channel State Information Reference Signal
CUDA	Compute Unified Device Architecture
cuBB	CUDA base-band (L1 software stack consisting of L2 adapter, PHY control layer and PHY layer)
CUDA	Compute Unified Device Architecture
cuPHY	CUDA PHY (L1 functionality on the GPU accelerator in inline mode)
DCI	Downlink Control Information
DL	Downlink
DMRS	Demodulation Reference Signal
DU or O-DU	O-RAN Distributed Unit (a logical node hosting RLC/MAC/High-PHY layers based on a lower layer functional split.)
eCPRI	Ethernet Common Public Radio Interface
eAxC	Extended Antenna Carrier: a data flow for a single antenna (or spatial stream) for a single carrier in a single sector
FAPI	Functional Application Programming Interface
FH	Fronthaul
H2D	Host-to-device memory
LDPC	Low-density Parity Check
NIC	Network interface card

O-RAN	Open RAN
PBCH	Physical Broadcast Channel
PDCCH	Physical Downlink Control Channel
PDSCH	Physical Downlink Shared Channel
PRACH	Physical Random Access Channel
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
RAN	Radio Access Network
RM	Reed-Muller
RU or O-RU	O-RAN Radio Unit: a logical node hosting Low-PHY layer and RF processing based on a lower layer functional split
SCF	Small Cell Forum
SSB	Synchronization Signal Block
SyncE	Synchronous Ethernet: is an ITU-T standard to provide a synchronization signal to network resources
UCI	Uplink Control Information
UL	Uplink
TB	Transport Block