



DOCA SDK Architecture

Table of contents

Device Subsystem

Memory Management Subsystem

Execution Model

DOCA provides libraries for networking and data processing programmability that leverage NVIDIA® BlueField® DPU and NVIDIA® ConnectX® NIC hardware accelerators.

DOCA software framework is built on top of DOCA Core, which provides a unified software framework for DOCA libraries, to form a processing pipeline or workflow build of one or many DOCA libraries.

Device Subsystem

The DOCA SDK allows applications to offload resource intensive tasks to HW, such as encryption, and compression.

The SDK also allows applications to offload network related tasks, such as packet acquisition, and RDMA send.

As such DPUs/NICs provide dedicated HW processing units for executing such tasks.

The DOCA device subsystem provides an abstraction of the HW processing units referred to as device.

DOCA Device subsystem provides means to:

- Discover available hardware acceleration units provided by DPUs/NICs
- Query capabilities and properties of available hardware acceleration units
- Open device to enable libraries to allocate and share resources necessary for HW acceleration

On a given system there can be multiple available devices. An application can choose a device based on the following characteristics:

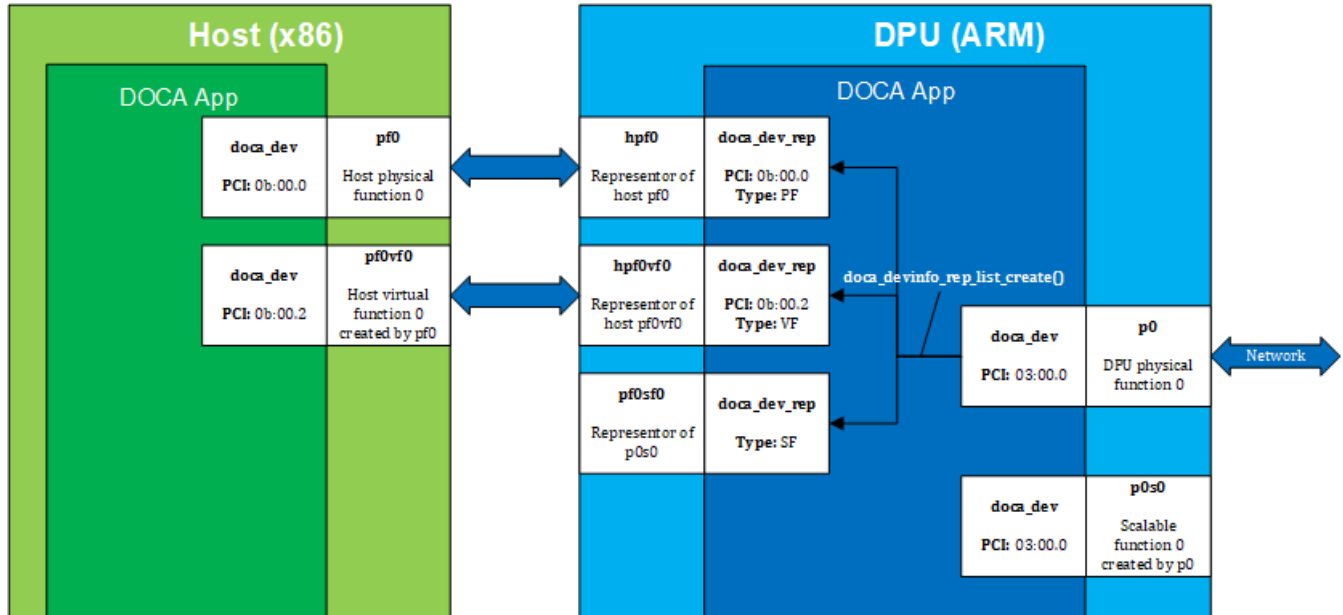
- Topology - E.g., PCIe address
- Capabilities - E.g., Encryption support

DOCA Core supports two DOCA Device types:

- Local device – this is an actual device exposed in the local system (DPU or host) and can perform DOCA library processing jobs. This can be a PCI physical function (PF) virtual function (VF) or scalable function (SF)

- Represor device – this is a representation of a local device. The represented local device is usually on the host (except for SFs) and the represor is always on the DPU side (a proxy on the DPU for the host-side device).

The following figure provides an example of host local devices with represors on DPU:



Note

The diagram shows typical topology when using a DPU in DPU mode as described in [NVIDIA BlueField DPU Modes of Operation](#).

The diagram shows a DPU (on the right side of the figure) connected to a host (on the left). The host has physical function PF0 with a child virtual function VF0.

The DPU side has a represor-device per each host function in a 1-to-1 relation (e.g., `hpf0` is the represor device for the host's PF0 device and so on) as well as a represor for each SF function, such that both the SF and its represor reside in the DPU.

For more details about DOCA Device subsystem, see section "[DOCA Device](#)".

Memory Management Subsystem

HW processing tasks require data buffers as inputs and/or outputs to processing operations. The application is responsible to provide the input data and/or read the output data.

In order to achieve maximum performance, the SDK uses zero-copy technology to pass data to the HW. To allow zero-copy, the application must register the memory that will hold data buffers beforehand.

The memory management subsystem provides a means to register memory and manage allocation of data buffers on registered memory.

Memory registration:

- Defines user application memory range that will be used to hold data buffers.
- Allows one or more devices to access the memory range.
- Defines the access permission (E.g., read only).

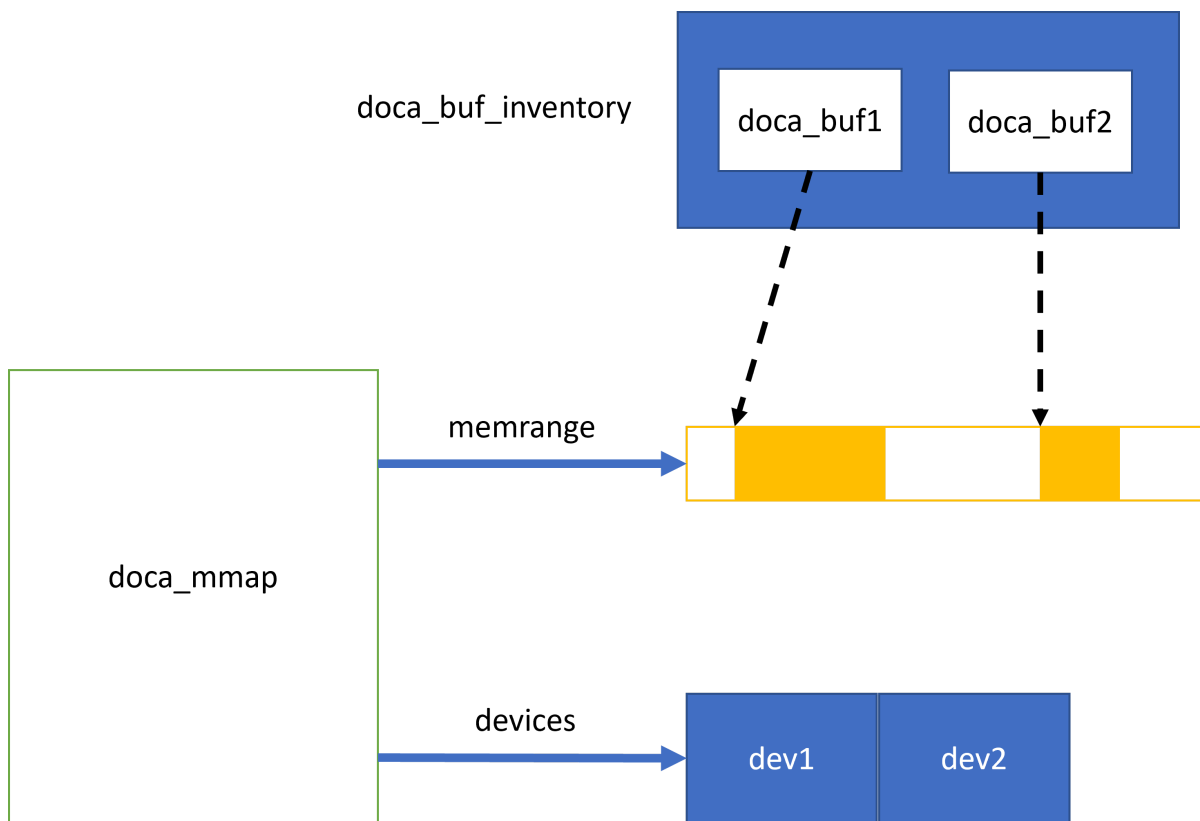
Data buffer allocation management:

- Allows allocating data buffers that cover subranges within the registered memory.
- Allows memory pool semantics over registered memory.

DOCA memory has the following main components:

- `doca_buf` – Describes a data buffer, and is used as input/output to various HW processing tasks within DOCA libraries.
- `doca_mmap` – Describes registered memory, that is accessible by devices, with a set of permissions. `doca_buf` is a segment in the memory range represented by `doca_mmap`.
- `doca_buf_inventory` – pool of `doca_buf` with the same characteristics (see more in sections "[DOCA Core Buffers](#)" and "[DOCA Core Inventories](#)")

The following diagram shows the various modules within the DOCA memory subsystem:



The diagram shows a `doca_buf_inventory` containing 2 `doca_buf`s. Each `doca_buf` points to a portion of the memory buffer which is part of a `doca_mmap`. The mmap is populated with one continuous memory range and is registered with Two DOCA Devices, dev1 and dev2.

For more details about DOCA Memory management subsystem, see section "[DOCA Memory Subsystem](#)".

Execution Model

DOCA SDK introduces libraries that utilize HW processing units. Each library defines dedicated APIs for achieving a specific processing task (E.g., Encryption). The library abstracts all the low level details related to operation of the HW, and instead lets the application focus on what matters. This type of library is referred to as a context.

Since a context utilizes a HW processing unit, it needs a device in order to operate. This device will also determine which buffers are accessible by that context.

Contexts provide HW processing operation APIs in the form of tasks and events.

Task:

- Application prepares the task arguments.

- Application submits the task, this will issue a request to the relevant HW processing unit.
- Application receives a completion in the form of a callback **once** the HW processing is completed.

Event:

- Application registers to the event. This will inform HW to report whenever the event occurs.
- Application receives a completion in form of a callback **every time** HW identifies that the event occurred.

Since HW processing is asynchronous in nature. DOCA provides an object that allows waiting on processing operations (tasks & events). This object is referred to as Progress Engine or PE.

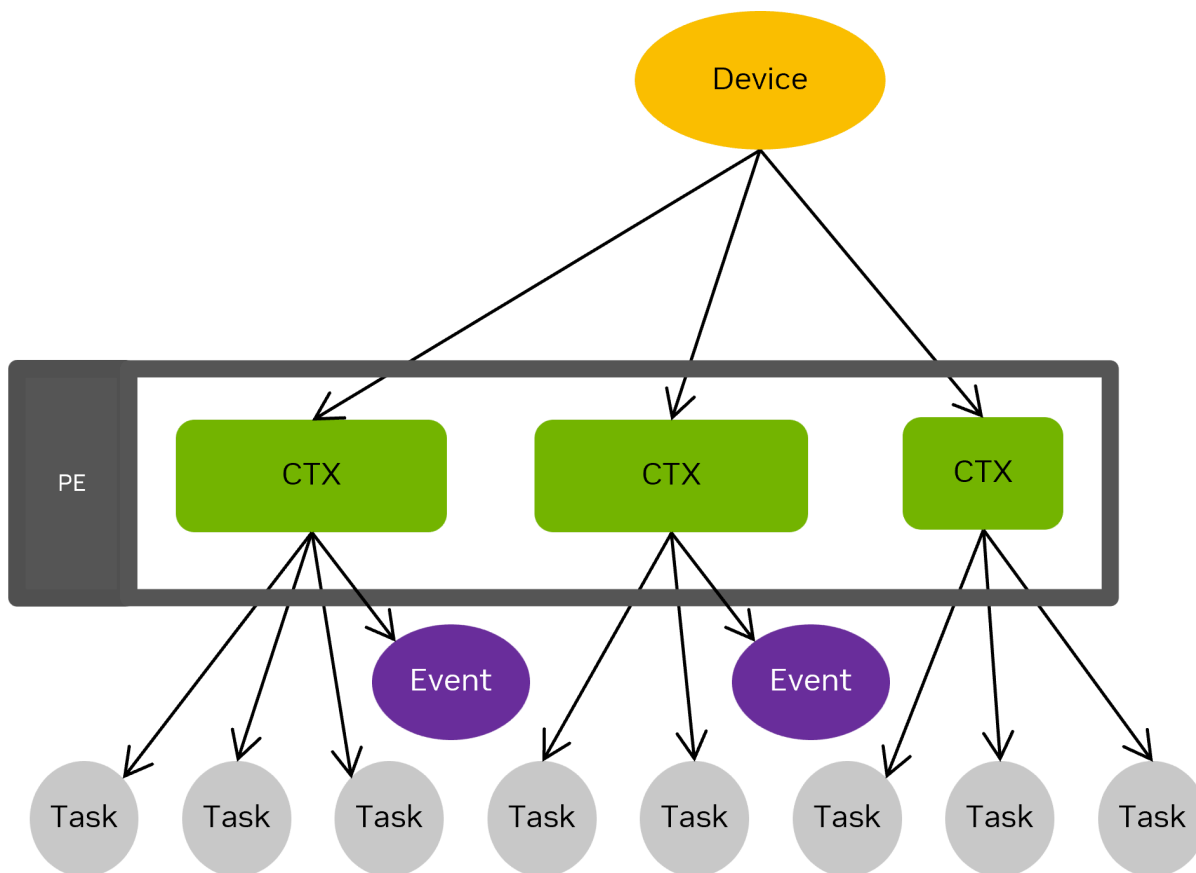
The PE allows waiting on completions using the following methods:

- Busy waiting/polling mode - in this case the application will repeatedly invoke a method that checks if some completion has occurred.
- Notification-driven mode - in this case the application can use OS primitives (E.g., linux event fd) to notify thread whenever some completion has occurred.

Once completion has occurred, whether caused by Task or Event, the relevant callback will be invoked as part of PE method.

A single PE instance allows waiting on multiple Tasks/Events from different contexts. As such it is possible for application to utilize a single PE per thread.

The following diagram shows how a combination of various DOCA modules combine DOCA cross-library processing runtime.



The diagram shows 3 contexts that are utilizing the same device, each context has some tasks/events that have been submitted/registered by application. All 3 contexts are connected to the same PE, where application can use same PE to wait on all completions at once.

For more details about DOCA Execution model see section "[DOCA Execution Model](#)".

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 and 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.

© Copyright 2024, NVIDIA. PDF Generated on 01/15/2025