NVIDIA DOCA Flow Inspector Service

Guide

MLNX-15-060530 _v1.4 | September 2022

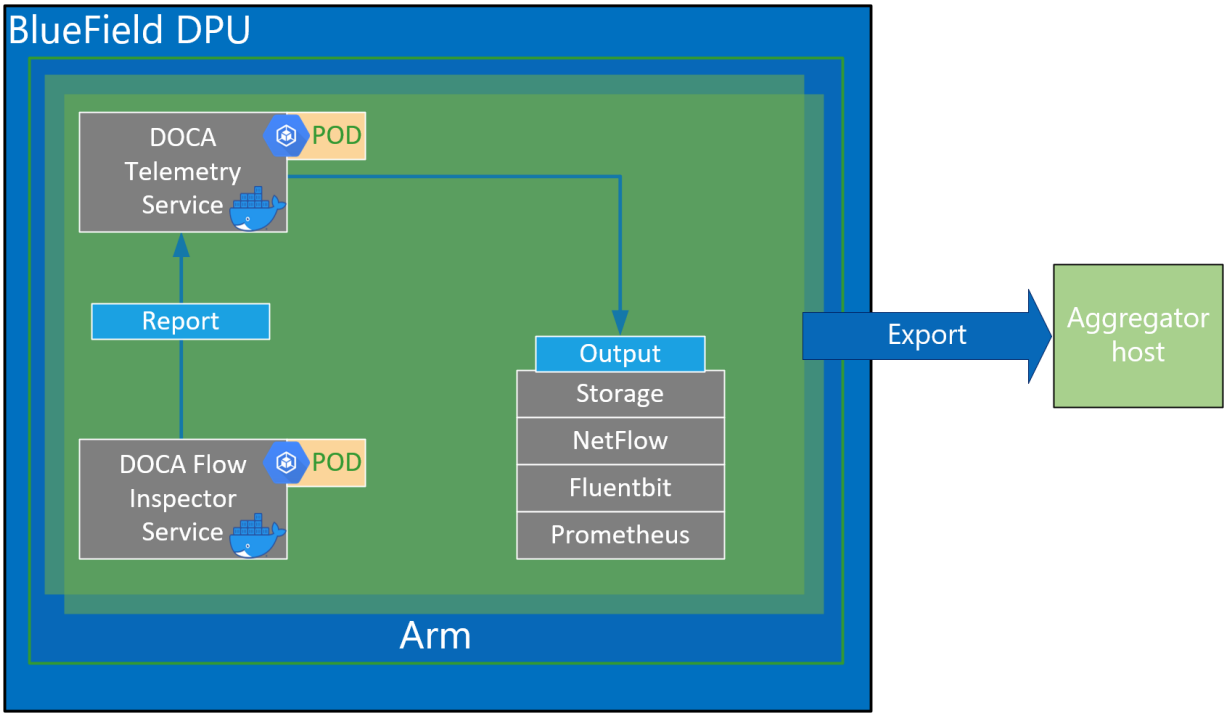
Table of Contents

Chapter 1. Introduction

DOCA Flow Inspector service allows monitoring real-time data and extraction of telemetry components which can be utilized by various services for security, big data, and other purposes.

DOCA Flow Inspector service is linked to the DOCA Telemetry Service (DTS). DOCA Flow Inspector receives mirrored packets from the user and then parses and forwards the data to the DTS which gathers predefined statistics forwarded by various providers/sources. DPCA Flow Inspector uses the DOCA Telemetry API to initiate a communication channel to the DTS while the DPDK infrastructure allows acquiring packets at a user-space layer.

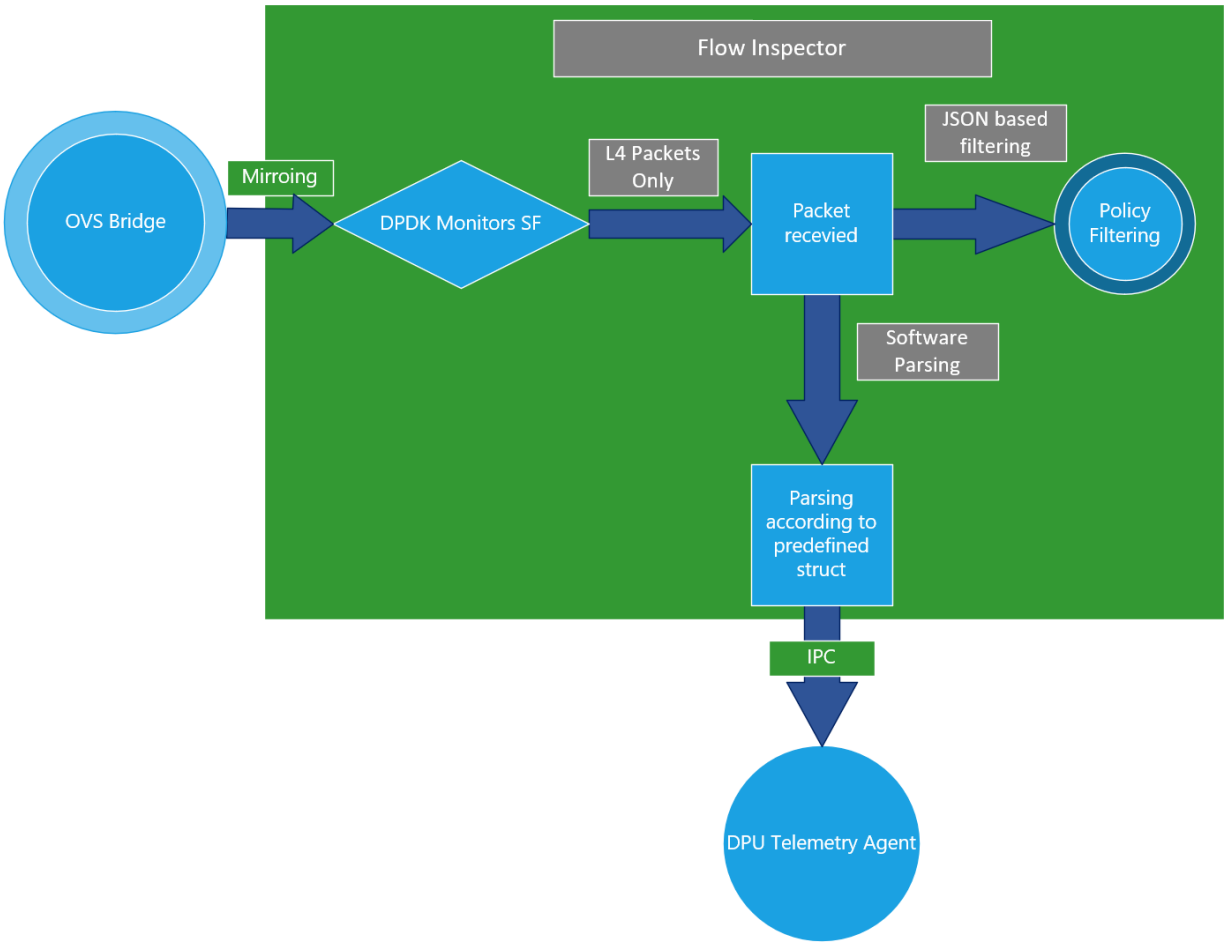
DOCA Flow Inspector runs inside of its own Kubernetes pod on BlueField and is intended to receive mirrored packets for analysis. The packets received are parsed and sent, in a predefined struct, to a telemetry collector which manages the rest of the telemetry aspects.



1.1. Service Flow

DOCA Flow Inspector receives a configuration file in a JSON format which indicates which of the mirrored packets should be filtered out based on the L4 network header.

DOCA Flow Inspector runs on top of DPDK to acquire L4. The packets are then filtered based on the ports configured in the JSON input file. The non-filtered packets are then parsed according to a predefined struct and forwarded to the telemetry collector using IPC.



1. A JSON file is used as input to configure the allowed ports.
2. Connection to the telemetry collector is initialized and a configuration struct is defined for both small and large packets.
3. Traffic is mirrored to the relevant SF.
4. Ingress traffic is received through the configured SF.
5. Non-L4 traffic is dropped in using hardware rules, while the allowed ports are filtered using a software layer in the Arm cores.
6. Packets are parsed to the desired struct.

7. The telemetry information is forwarded to the telemetry agent using IPC.
8. Mirrored packets are freed.

The information collected by DOCA Flow Inspector has two different structures that are differentiated by the payload's size (can be viewed when using the DTS to write the date locally).

The predefined struct is organized as follows:

1. Timestamp
2. Host IP
3. Data Length
4. Data (Payload)
5. Source MAC
6. Destination MAC
7. L4 Protocol
8. Source IP
9. Destination IP
10. Source Port
11. Destination Port
12. TCP Flags

Chapter 2. Requirements

DOCA Flow Inspector service must be used with DTS of the same DOCA version.

Before deploying the flow inspector container, ensure that the following prerequisites are satisfied:

1. Create the needed files and directories.

   Folders should be created automatically. Make sure the `.json` file resides inside the folder:

   ```
   $ touch /opt/mellanox/doca/services/flow_inspector/bin/flow_inspector_cfg.json
   ```

   Validate that DTS's configuration folders exist. They should be created automatically when deploying DTS using the `.yaml` file.

   ```
   $ sudo mkdir -p /opt/mellanox/doca/services/telemetry/config
   $ sudo mkdir -p /opt/mellanox/doca/services/telemetry/ipc_sockets
   $ sudo mkdir -p /opt/mellanox/doca/services/telemetry/data
   ```

2. Allocate hugepages.

   Allocated huge pages for a DPDK-based application. This requires root privileges.

   ```
   $ sudo echo 2048 > /sys/kernel/mm/hugepages/hugepages-2048kB/nr_hugepages
   ```

   Or alternatively:

   ```
   $ sudo echo '2048' | sudo tee -a /sys/kernel/mm/hugepages/hugepages-2048kB/
   nr_hugepages
   $ sudo mkdir /mnt/huge
   $ sudo mount -t hugetlbfs nodev /mnt/huge
   ```

   Deploy a scalable function according to Scalable Function Setup Guide and mirror packets accordingly using the Open vSwitch command.

   For example:

   a). Mirror packets from `p0` to `sf0`:

   ```
   $ ovs-vsctl add-br ovsbr1
   $ ovs-vsctl add-port ovsbr1 p0
   $ ovs-vsctl add-port ovsbr1 en3f0pf0sf0
   $ ovs-vsctl -- --id=@p1 get port en3f0pf0sf0 \
               -- --id=@p2 get port p0 \
               -- --id=@m create mirror name=m0 select-dst-port=@p2 select-src-
   port=@p2 output-port=@p1 \
               -- set bridge ovsbr1 mirrors=@m
   ```

   b). Mirror packets from `pf0hpf` to `sf0`:

   ```
   $ ovs-vsctl add-br ovsbr1
   $ ovs-vsctl add-port ovsbr1 pf0hpf
   $ ovs-vsctl add-port ovsbr1 en3f0pf0sf0
   $ ovs-vsctl -- --id=@p1 get port en3f0pf0sf0 \
               -- --id=@p2 get port pf0hpf \
   ```

```
            -- --id=@m create mirror name=m0 select-dst-port=@p2 select-src-
port=@p2 output-port=@p1 \
            -- set bridge ovsbr1 mirrors=@m
```

The output of last command should be in the following format:

```
exp: 0d248ca8-66af-427c-b600-af1e286056e1
```

Note: The designated SF must be created as a trusted function. Additional details can be found in the Scalable Function Setup Guide.

Chapter 3. Service Deployment

For information about the deployment of DOCA containers on top of the BlueField DPU, refer to NVIDIA DOCA Container Deployment Guide.

DTS is available on NGC, NVIDIA's container catalog. Service-specific configuration steps and deployment instructions can be found under the service's container page.

Chapter 4. Configuration

4.1. JSON Input

The flow inspector configuration file should be placed under `/opt/mellanox/services/flow_inspector/<json_file_name>.json` and be built in the following format:

```
{
    "protocols": ["tcp", "udp"], #What L4 protocols are allowed
    "ports": #What L4 port ranges are allowed [src,dst]
    [
        ["*","433"], #In this case, "*" stands for all source ports and dst port 433
        ["20480","28341"],
        ["28341","20480"],
        ["1720","1023"], # src port 1720, dst port 1023
    ]
}
```

Note: If a packet contains L4 ports which are not specified in the file, it is filtered out.

4.2. Yaml File

The `.yaml` file downloaded from NGC can be easily edited according to your needs.

```
env:
  # Set according to the local setup
  - name: SF_NUM_1
    value: "2"
  # Additional EAL flags, if needed
  - name: EAL_FLAGS
    value: ""
  # Service-Specific command line arguments
  - name: SERVICE_ARGS
    value: "--policy /flow_inspector/flow_inspector_cfg.json -l 4"
```

- The `SF_NUM_1` value can be changed according to the SF used in the OVS configuration and can be found using the command in Scalable Function Setup Guide.
- The `EAL_FLAGS` value must be changed according to the DPDK flags required when running the container.
- The `SERVICE_ARGS` are the runtime arguments received by the service:
  - `-l, --log-level <value>` – sets the log level from 0 to 4

- `-p, --policy <json_path>` – sets the JSON path inside the container

4.3. Verifying Output

Enabling write to data in the DTS allows debugging the validity of DOCA Flow Inspector.

The data written locally should be shown in the following format, which has been parsed by the flow inspector:

```
doca_telemetry_long_event {
    doca_telemetry_timestamp_t timestamp;
    char host_ip[MAX_IP_LEN];
    int64_t data_len;
    char data[MAX_DATA_LEN_LONG];
    char src_mac[MAX_MAC_LEN];
    char dest_mac[MAX_MAC_LEN];
    char protocol[PROTOCOL_LEN];
    char src_ip[MAX_IP_LEN];
    char dest_ip[MAX_IP_LEN];
    int src_port;
    int dest_port;
    int flags;
};
```

Uncomment the following line in `dts_config.ini`:

```
#output=/data
```

Note: Any changes in `dts_config.ini` necessitate restarting the pod for the new settings to apply.

The schema folder contains JSON-formatted metadata files which allow reading the binary files containing the actual data. The binary files are written according to the naming convention shown in the following example (`apt install tree`):

```
$ tree /opt/mellanox/doca/services/telemetry/data/
/opt/mellanox/doca/services/telemetry/data/
├── {year}
│   └── {mmdd}
│       └── {hash}
│           ├── {source_id}
│           │   └── {source_tag}{timestamp}.bin
│           └── {another_source_id}
│               └── {another_source_tag}{timestamp}.bin
└── schema
    └── schema_{MD5_digest}.json
```

New binary files appear when the service starts or when binary file's max age/size restriction is reached. If no schema or no data folders are present, refer to the Troubleshooting section in the NVIDIA DOCA Telemetry Service Guide.

Note: `source_id` is usually set to the machine hostname. `source_tag` is a line describing the collected counters, and it is often set as the provider's name or name of user-counters.

Reading the binary data can be done from within the DTS container using the following command:

```
crictl exec -it <Container ID> /opt/mellanox/collectx/bin/clx_read -s /data/schema /
data/path/to/datafile.bin
```

Chapter 5. Troubleshooting

Refer to the Troubleshootings section in NVIDIA DOCA Container Deployment Guide and NVIDIA DOCA Telemetry Service Guide.

Additional notes:

- Make sure hugepages are allocated (step 2 under Requirements)
- Validate the JSON file and port config

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 nor any of its direct or indirect subsidiaries and affiliates (collectively: "NVIDIA") make no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assume 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, and Mellanox are trademarks and/or registered trademarks of Mellanox Technologies Ltd. and/or NVIDIA Corporation in the U.S. and in other countries. The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a world¬wide basis. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

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