NVIDIA DOCA Flow

Programming Guide

MLNX-15-060488 _v1.3 | June 2022

Table of Contents

Chapter 1. Introduction

DOCA Flow is the most fundamental API for building generic execution pipes in hardware.

The library provides an API for building a set of pipes, where each pipe consists of match criteria, monitoring, and a set of actions. Pipes can be chained so that after a pipe-defined action is executed, the packet may proceed to another pipe.

Using DOCA Flow API, it is easy to develop HW-accelerated applications that have a match on up to two layers of packets (tunneled).

- MAC/VLAN/ETHERTYPE
- IPv4/IPv6
- TCP/UDP/ICMP
- GRE/VXLAN/GTP-U
- Metadata

The execution pipe may include packet modification actions:

- Modify MAC address
- Modify IP address
- Modify L4 (ports, TCP sequences, and acknowledgments)
- Strip tunnel
- Add tunnel
- Set metadata

The execution pipe may also have monitoring actions:

- Count
- Policers
- Mirror

The pipe also has a forwarding target which may be any of the following:

- Software (RSS to subset of queues)
- Port
- Another pipe

- Drop packets

This document is intended for software developers writing network function applications that focus on packet processing (e.g., gateways). The document assumes familiarity with network stack and DPDK.

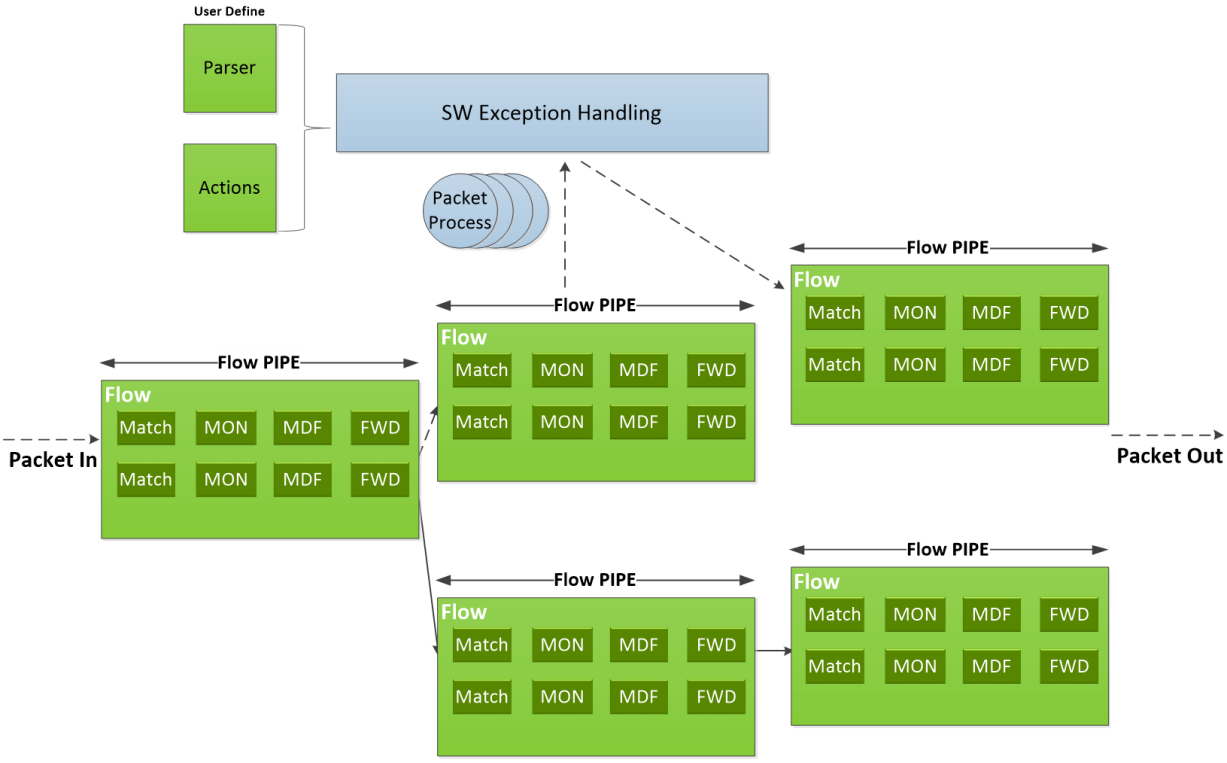
Chapter 2. Prerequisites

A DOCA Flow-based application can run either on the host machine or on the NVIDIA® BlueField® DPU target. Since it is based on DPDK, Flow-based programs require an allocation of huge pages:

```
echo 1024 > /sys/kernel/mm/hugepages/hugepages-2048kB/nr_hugepages
sudo mkdir /mnt/huge
sudo mount -t hugetlbfs nodev /mnt/huge
```

Chapter 3. Architecture

The following diagram shows how the DOCA Flow library defines a pipe template, receives a packet for processing, creates the a pipe entry, and offloads the flow rule in HW NIC.



- MON: Monitor, can be count, meter, or mirror. MDF: Modify function, can be modify a field. FWD: Forward to next stage in packet processing.
- User-defined set of matches parser and actions.
- DOCA Flow pipes can be created or destroyed dynamically.
- Packet processing is fully accelerated by hardware with a specific entry in flow pipe.
- Packets that do not match any of the pipe entries in hardware can be sent to Arm cores for exception handling and then reinjected back to hardware.

Chapter 4. API

Refer to NVIDIA DOCA Libraries API Reference Manual, for more detailed information on DOCA Flow API.

Note: The pkg-config (`*.pc` file) for the telemetry library is named `doca-flow`.

The following sections provide additional details about the library API.

4.1. doca_flow_cfg

This is the DOCA Flow initialize configuration struct.

```
struct doca_flow_cfg {
 uint16_t queues;
 struct doca_flow_resources resource;
 const char *mode_args;
 bool aging;
 uint32_t nr_shared_resources[DOCA_FLOW_SHARED_RESOURCE_MAX];
 unit32_t queue_depth;
 doca_flow_entry_process_cb cb;
};
```

queues
The number of hardware acceleration control queues. It is expected that the same core always uses the same `queue_id`. In cases where multiple cores access the API using the same `queue_id`, it is up to the application to use locks between different cores/threads.

resource
Resource quota. This field includes the flow resource quota defined in the following structs:

- `uint32_t nb_counters` – number of counters to configure
- `uint32_b nb_meters` – number of traffic meters to configure

mode_args
Set the DOCA Flow architecture mode.

aging
Aging is handled by DOCA Flow while it is set to true.

nr_shared_resources
Total shared resource per type.

queue_depth
Number of pre-configured `queue_size`. Default is 128.

cb
Callback function for entry create/destroy

4.2. doca_flow_port_cfg

This is the DOCA Flow initialize port configuration struct.

```
struct doca_flow_port_cfg {
 uint16_t port_id;
 enum doca_flow_port_type type;
 const char *devargs;
 uint16_t priv_data_size;
};
```

port_id
DPDK port ID.

type
Depends on underlying API. This field includes the following port type:

- `DOCA_FLOW_PORT_DPDK_BY_ID` – DPDK port by mapping ID

devargs
String containing the exact configuration needed according to the type.

> **Note:** For usage information of the type and devargs fields, refer to Start Port.

priv_data_size
Per port, users may define private data where application-specific information can be stored.

4.3. doca_flow_pipe_cfg

This is a pipe configuration that contains the user-defined template for the packet process. The library is configured according to the struct fields.

```
struct doca_flow_pipe_cfg {
        const char *name;
 enum doca_flow_pipe_type type;
 struct doca_flow_port *port;
 bool is_root;
 struct doca_flow_match *match;
 struct doca_flow_match *match_mask;
 struct doca_flow_actions *actions;
 struct doca_flow_action_descs *action_descs;
 struct doca_flow_monitor *monitor;
};
```

name
Name for the pipeline

type
Type of pipe (`enum doca_flow_pipe_type`). This field includes the following pipe types:

- `DOCA_FLOW_PIPE_BASIC` – flow pipe
- `DOCA_FLOW_PIPE_CONTROL` – control pipe

port
Port for the pipeline. Refer to `doca_flow_port_start`.

is_root
Determins whether pipeline is root or not.

match
Matcher for the pipeline. Refer to `doca_flow_match`.

match_mask
Match mask for the pipeline. Refer to `doca_flow_match`.

actions
Actions for the pipeline. Refer to `doca_flow_actions`.

action_descs
Action descriptions. Refer to `doca_flow_action_desc`.

monitor
Monitor for the pipeline. Refer to `doca_flow_monitor`.

4.4. doca_flow_meta

This is a maximum 20-byte scratch area which exists throughout the pipeline.

The user can set a value to metadata, copy from a packet field, then match in later pipes. Mask is supported in both match and modification actions.

The user can modify the metadata in different ways based on its description type:

AUTO
Set metadata value from action of a specific entry. Pipe action is used as mask.

CONSTANT
Set metadata value from pipe action. Masked by description mask.

SET
Set metadata value from action of a specific entry. Masked by description as mask.

Note: In a real application, it is encouraged to create a union of `doca_flow_meta` defining the application's scratch fields to use as metadata.

```
struct doca_flow_meta {
    uint32_t pkt_meta;
    uint32_t u32[];
}
```

pkt_meta
Metadata can be received along with packet.

u32[]
Scratch area.

Note: If `encap` action is used, `pkt_meta` should not be defined by the user as it is defined internally in DOCA to reference the encapsulated tunnel ID.

4.5. doca_flow_match

This function is a match configuration that contains the user-defined fields that should be matched on the pipe.

```
struct doca_flow_match {
 uint32_t flags;
 struct doca_flow_meta meta;
 uint8_t out_src_mac[DOCA_ETHER_ADDR_LEN];
 uint8_t out_dst_mac[DOCA_ETHER_ADDR_LEN];
 doca_be16_t out_eth_type;
 doca_be16_t out_vlan_id;
 struct doca_flow_ip_addr out_src_ip;
 struct doca_flow_ip_addr out_dst_ip;
 uint8_t out_l4_type;
 uint8_t out_tcp_flags;
 doca_be16_t out_src_port;
 doca_be16_t out_dst_port;
 struct doca_flow_tun tun;
 uint8_t in_src_mac[DOCA_ETHER_ADDR_LEN];
 uint8_t in_dst_mac[DOCA_ETHER_ADDR_LEN];
 doca_be16_t in_eth_type;
 doca_be16_t in_vlan_id;
 struct doca_flow_ip_addr in_src_ip;
 struct doca_flow_ip_addr in_dst_ip;
 uint8_t in_l4_type;
 uint8_t in_tcp_flags;
 doca_be16_t in_src_port;
 doca_be16_t in_dst_port;
};
```

flags
Match items which are no value needed.

meta
Programmable meta data.

out_src_mac
Outer source MAC address.

out_dst_mac
Outer destination MAC address.

out_eth_type
Outer Ethernet layer type.

out_vlan_id
Outer VLAN ID.

out_src_ip
Outer source IP address.

out_dst_ip
Outer destination IP address.

out_l4_type
Outer layer 4 protocol type.

out_tcp_flags
Outer TCP flags.

out_src_port
Outer layer 4 source port.

out_dst_port
Outer layer 4 destination port.

tun
Tunnel info.

in_src_mac
Inner source MAC address if tunnel is used.

in_dst_mac
Inner destination MAC address if tunnel is used.

in_eth_type
Inner Ethernet layer type if tunnel is used.

in_vlan_id
Inner VLAN ID if tunnel is used.

in_src_ip
Inner source IP address if tunnel is used.

in_dst_ip
Inner destination IP address if tunnel is used.

in_l4_type
Inner layer 4 protocol type if tunnel is used.

in_tcp_flags
Inner TCP flags if tunnel is used.

in_src_port
Inner layer 4 source port if tunnel is used.

in_dst_port
Inner layer 4 destination port if tunnel is used.

4.6. doca_flow_actions

This function is a flow actions configuration.

```
struct doca_flow_actions {
 bool decap;
 uint8_t mod_src_mac[DOCA_ETHER_ADDR_LEN];
 uint8_t mod_dst_mac[DOCA_ETHER_ADDR_LEN];
 struct doca_flow_ip_addr mod_src_ip;
 struct doca_flow_ip_addr mod_dst_ip;
 doca_be16_t mod_src_port;
 doca_be16_t mod_dst_port;
 bool dec_ttl;
 bool has_encap;
 struct doca_flow_encap_action encap;
 struct doca_flow_meta meta;
};
```

decap
Decap while it is set to true.

mod_src_mac
Modify source MAC address.

mod_dst_mac
Modify destination MAC address.

mod_src_ip
Modify source IP address.

mod_dst_ip
Modify destination IP address.

mod_src_port
Modify layer 4 source port.

mod_dst_port
Modify layer 4 destination port.

dec_ttl
Decrease TTL value while it is set to true.

has_encap
Encap while it is set to true.

encap
Encap data information.

meta
Mask of metadata if action description type is AUTO, constant value if type is CONSTANT.

4.7. doca_flow_action_desc

This function is an action description.

```
struct doca_flow_action_desc {
 enum doca_flow_action_type type;
 union {
  union {
   struct doca_flow_meta meta;
   uint64_t u64;
    uint8_t u8[16];
  } mask;
  struct {
   unit16_t doca_flow_action_field src;
   unit16_t doca_flow_action_field dst;
   unit16_t width;
  } copy;
  struct { The type field includes the forwarding modification types
   uint64_t val;
   struct doca_flow_action_field dst;
  } add;
 };
};
```

type
Action type.

mask
Mask of modification type CONSTANT and SET.

copy
Field copy source and destination description.

add
Field add description.

The `type` field includes the following forwarding modification types:

- `DOCA_FLOW_ACTION_AUTO` – modification type derived from pipe action
- `DOCA_FLOW_ACTION_CONSTANT` – modify with the constant value from pipe
- `DOCA_FLOW_ACTION_SET` – modify field with the value of pipe entry
- `DOCA_FLOW_ACTION_ADD` – add field value. Supports `ipv4_ttl`, `ipv6_hop`, `tcp_seq`, and `tcp_ack`.

- DOCA_FLOW_ACTION_COPY – copy field.

Refer to Setting Pipe Actions for more information.

4.8. doca_flow_monitor

This function is a monitor configuration.

```
struct doca_flow_monitor {
 uint8_t flags;
 struct {
  unit32_t cir;
  unit32_t cbs;
 };
 uint32_t aging;

 uint32_t user_data;
};
```

flags
Indicate actions to be included.

cir
Committed information rate in bytes per second. Defines maximum bandwidth.

cbs
Committed burst size in bytes. Defines maximum local burst size.

aging
Aging time in seconds.

user_data
Aging user data input.

The flags field includes the following monitor types:

- DOCA_FLOW_ACTION_METER – set monitor with meter action
- DOCA_FLOW_ACTION_COUNT – set monitor with counter action
- DOCA_FLOW_ACTION_AGING – set monitor with aging action

T(c) is the number of available tokens. For each packet where b equals the number of bytes, if t(c)−b≥0 the packet can continue, and tokens are consumed so that t(c)=t(c)−b. If t(c)−b<0, the packet is dropped.

T(c) tokens are increased according to time, configured CIR, configured CBS, and packet arrival. When a packet is received, prior to anything else, the t(c) tokens are filled. The number of tokens is a relative value that relies on the total time passed since the last update, but it is limited by the CBS value.

CIR is the maximum bandwidth at which packets continue being confirmed. Packets surpassing this bandwidth are dropped. CBS is the maximum bytes allowed to exceed the CIR to be still CIR confirmed. Confirmed packets are handled based on the fwd parameter.

The number of <cir,cbs> pair different combinations is limited to 128.

4.9. doca_flow_fwd

This function is a forward configuration which directs where the packet goes next.

```
struct doca_flow_fwd {
 enum doca_flow_fwd_type type;
 union {
  struct {
   unit32_t rss_flags;
   unit32_t *rss_queues;
   int num_of_queues;
   uint32_t rss_mark;
  };
  struct {
   unit16_t port_id;
  };
  struct {
   struct doca_flow_pipe *next_pipe;
  };
 };
};
```

type
Indicates the forwarding type.

rss_flags
RSS offload types.

rss_queues
RSS queues array.

num_of_queues
Number of queues.

rss_mask
Mark ID of each queue.

port_id
Destination port ID.

next_pipe
Next pipe pointer.

The `type` field includes the forwarding action types defined in the following enum:

- `DOCA_FLOW_FWD_RSS` – forwards packets to RSS
- `DOCA_FLOW_FWD_PORT` – forwards packets to port
- `DOCA_FLOW_FWD_PIPE` – forwards packets to another pipe
- `DOCA_FLOW_FWD_DROP` – drops packets

The `rss_flags` field includes the RSS fields defined in the following enum:

- `DOCA_FLOW_RSS_IP` – RSS by IP header
- `DOCA_FLOW_RSS_UDP` – RSS by UDP header
- `DOCA_FLOW_RSS_TCP` – RSS by TCP header

4.10. doca_flow_query

This struct is a flow query result.

```
struct doca_flow_query {
 uint64_t total_bytes;
 uint64_t total_pkts;
};
```

total_bytes
Total bytes hit this flow.

total_pkts
Total packets hit this flow.

4.11. doca_flow_aged_query

This function is an aged flow callback context.

```
struct doca_flow_aged_query {
 uint64_t user_data;
};
```

user_data
The user input context. Otherwise, the `doca_flow_pipe_entry` pointer be returned.

4.12. doca_flow_init

This function is the global initialization function for DOCA Flow.

```
int doca_flow_init(const struct doca_flow_cfg *cfg, struct doca_flow_error *error);
```

cfg [in]
A pointer to flow config structure.

error [out]
A pointer to flow error output.

Returns
0 on success, a negative `errno` value otherwise and error is set.

Note: Must be invoked first before any other function in this API. This is a one-time call used for DOCA Flow initialization and global configurations.

4.13. doca_flow_port_start

This function starts a port with its given configuration. It creates one port in the DOCA Flow layer, allocates all resources used by this port, and creates the default offload flows including jump and default RSS for traffic.

```
struct doca_flow_port *doca_flow_port_start(const struct doca_flow_port_cfg *cfg,
                                            struct doca_flow_error *error);
```

cfg [in]
A pointer to flow port config structure.

error [out]
A pointer to flow error output.

Returns
Port handler on success, NULL otherwise an error is set.

4.14. doca_flow_port_pair

This function pairs two DOCA ports. If two ports are not representor ports, after performing a physical hairpin bind, this API notifies DOCA that these two ports are hairpin peers. If FWD to the hairpin port, DOCA builds a hairpin queue action. If one of the two ports is a representor, DOCA creates a miss flow with a port action to redirect the traffic from one port to the other. Those two paired ports have no order, and a port cannot be paired with itself.

```
int *doca_flow_port_pair(struct doca_flow_port *port,
                         struct doca_flow_port *pair_port);
```

port [in]
A pointer to DOCA Flow port structure.

pair_port [in]
A pointer to another DOCA Flow port structure.

Returns
0 on success, negative value on failure.

4.15. doca_flow_create_pipe

This function creates a new pipeline to match and offload specific packets. The pipeline configuration is defined in the `doca_flow_pipe_cfg`. The API creates a new pipe but does not start the hardware offload.

When `cfg` type is `DOCA_FLOW_PIPE_CONTROL`, the function creates a special type of pipe that can have dynamic matches and forwards with priority. The number of entries is limited to <64.

```
struct doca_flow_pipe *
doca_flow_create_pipe(const struct doca_flow_pipe_cfg *cfg,
                      const struct doca_flow_fwd *fwd,
                      const struct doca_flow_fwd *fwd_miss,
                      struct doca_flow_error *error);
```

cfg [in]
A pointer to flow pipe config structure.

fwd [in]
A pointer to flow forward config structure.

fwd_miss [in]
A pointer to flow forward miss config structure. NULL for no `fwd_miss`. When creating a pipe, if there is a miss and `fwd_miss` is configured, then packet steering should jump to it.

error [out]
A pointer to flow error output.

Returns
Pipe handler on success, NULL otherwise and error is set.

4.16. doca_flow_pipe_add_entry

This function add a new entry to a pipe. When a packet matches a single pipe, it starts hardware offload. The pipe defines which fields to match. This API does the actual hardware offload, with the information from the fields of the input packets.

```
struct doca_flow_pipe_entry *
doca_flow_pipe_add_entry(uint16_t pipe_queue,
                         struct doca_flow_pipe *pipe,
                         const struct doca_flow_match *match,
                         const struct doca_flow_actions *actions,
                         const struct doca_flow_monitor *monitor,
                         const struct doca_flow_fwd *fwd,
   unit32_t flags,
   void *usr_ctx,
                         struct doca_flow_error *error);
```

pipe_queue [in]
Queue identifier.

pipe [in]
A pointer to flow pipe.

match [in]
A pointer to flow match. Indicates specific packet match information.

actions [in]
A pointer to flow modify actions. Iindicates specific packet modify information.

monitor [in]
A pointer to flow meter profiling or aging.

fwd [in]
A pointer to flow forward actions.

flags [in]
Determines whether flow entry is pushed to hardware immediately or not.

usr_ctx [in]
A pointer to user context.

error [out]
A pointer to flow error output.

Returns
Pipe entry handler on success, NULL otherwise and error is set.

The `flags` parameter includes the flow `flags` type defined in the following enum:

- `DOCA_FLOW_NO_WAIT` – the entry is not buffered
- `DOCA_FLOW_WAIT_FOR_BATCH` – the entry is buffered

4.17. doca_flow_control_pipe_add_entry

This function adds one new entry to a control pipe. The maximum number of entries per pipe is limited to <64.

```
struct doca_flow_pipe_entry *
doca_flow_control_pipe_add_entry(uint16_t pipe_queue,
                         uint8_t priority,
```

```
                            struct doca_flow_pipe *pipe,
                            const struct doca_flow_match *match,
                            const struct doca_flow_match *match_mask,
                            const struct doca_flow_fwd *fwd,
                            struct doca_flow_error *error);
```

pipe_queue [in]
Queue identifier.

priority [in]
Priority value.

pipe [in]
A pointer to flow pipe.

match [in]
A pointer to flow match. Indicates specific packet match information.

match_mask [in]
A pointer to flow match mask information.

fwd [in]
A pointer to flow FWD actions.

error [out]
A pointer to flow error output.

Returns
Pipe entry handler on success, NULL otherwise and error is set.

Refer to Miss Pipe and Control Pipe for more information.

4.18. doca_flow_entries_process

This function processes entries in the queue. The application must invoke this function to complete flow rule offloading and to receive the flow rule's operation status.

```
int
doca_flow_entries_process(struct doca_flow_port *port,
                          uint16_t pipe_queue,
                          uint64_t timeout,
                          uint32_t max_processed_entries);
```

port [in]
A pointer to the flow port structure.

pipe_queue [in]
Queue identifier.

timeout [in]
Priority value.

max_processed_entries [in]
A pointer to the flow pipe.

Returns
>0 – the number of entries processed
0 – no entries are processed
<0 – failure

4.19. doca_flow_entries_process

This function get the status of pipe entry.

```
enum doca_flow_entry_status
doca_flow_entry_get_status(struct doca_flow_entry *entry);
```

entry [in]
A pointer to the flow pipe entry to query.

Returns
Entry's status, defined in the following enum:

- DOCA_FLOW_ENTRY_STATUS_IN_PROCESS – the operation is in progress
- DOCA_FLOW_ENTRY_STATUS_SUCCESS – the operation completed successfully
- DOCA_FLOW_ENTRY_STATUS_ERROR – the operation failed

4.20. doca_flow_query

This function queries the packet statistics about a specific pipe entry.

```
int doca_flow_query(struct doca_flow_pipe_entry *entry, struct doca_flow_query
 *query_stats);
```

entry [in]
A pointer to the flow pipe entry to query.

query_stats [out]
A pointer to the data retrieved by the query.

Returns
0 on success, a negative errno value otherwise and error is set.

4.21. doca_flow_handle_aging

This function handles aging of flows in a queue. It goes over all flows and releases aged flows from being tracked. The entries array is filled with aged flows. Since the number of flows can be very large, it can take a significant amount of time to go over all flows, so this function is limited by a time quota. This means it might return without handling all flows which requires the user to call it again.

```
int doca_flow_query(struct doca_flow_pipe_entry *entry, struct doca_flow_query
 *query_stats);
```

queue [in]
Queue identifier.

quota [in]
Max time quota in microseconds for this function to handle aging.

entries [in]
User input entry array for the aged flows.

len [in]
User input length of entries array.

Returns

>0 – the number of aged flows filled in entries array.

0 – no aged entries in current call.

-1 – full cycle is done.

Chapter 5. Flow Life Cycle

5.1. Initialization Flow

Before using any DOCA Flow function, it is mandatory to call DOCA Flow initialization, `doca_flow_init()`, which initializes all resources used by DOCA Flow.

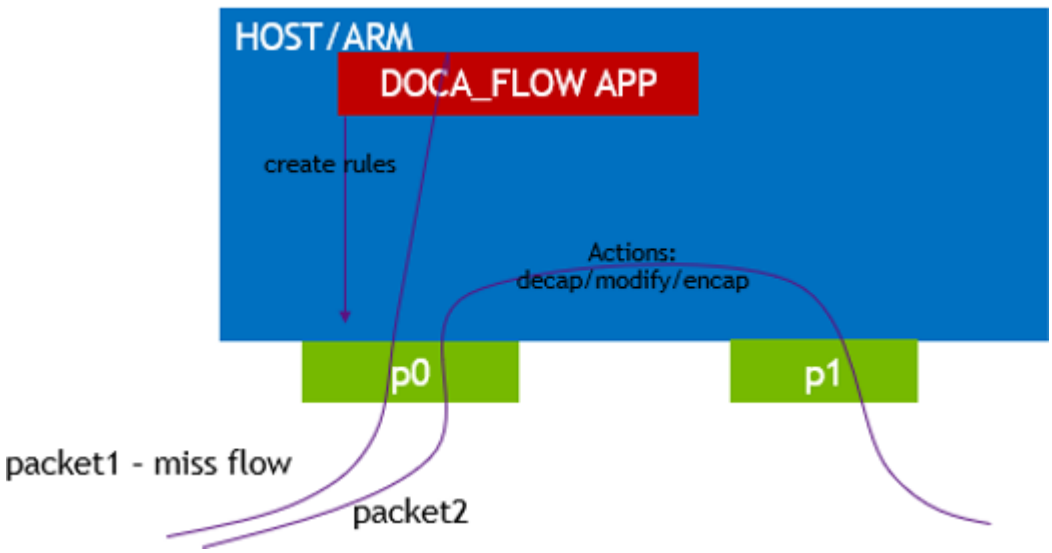
5.1.1. Pipe Mode

This mode defines the basic traffic in DOCA. It creates some miss rules when the DOCA port initialized. Currently, DOCA supports 3 types:

- `vnf`

  The packet arrives from one side of the application, is processed, and sent from the other side. The miss packet by default goes to the RSS of all queues.

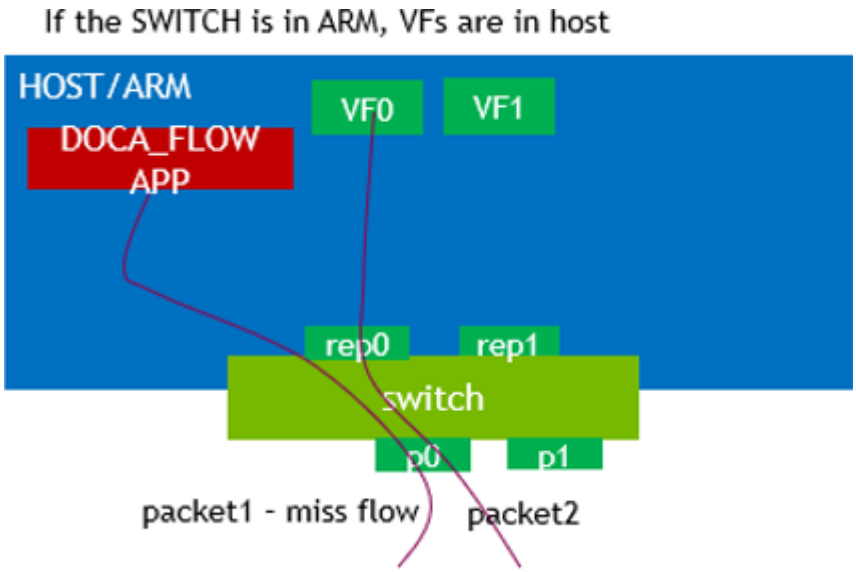
  The following diagram shows the basic traffic flow in `vnf` mode. Packet1 firstly misses to host RSS queues. The app captures this packet and decides how to process it and then creates a pipe entry. Packet2 will hit this pipe entry and do the action, for example, for VXLAN, will do decap, modify, and encap, then is sent out from P1.



- `switch`

  Used for internal switching, only representor ports are allowed, for example, uplink representors and SF/VF representors. Packet is forwarded from one port to another. If a packet arrives from an uplink and does not hit the rules defined by the user's pipe. Then the packet is received on all RSS queues of the representor of the uplink.

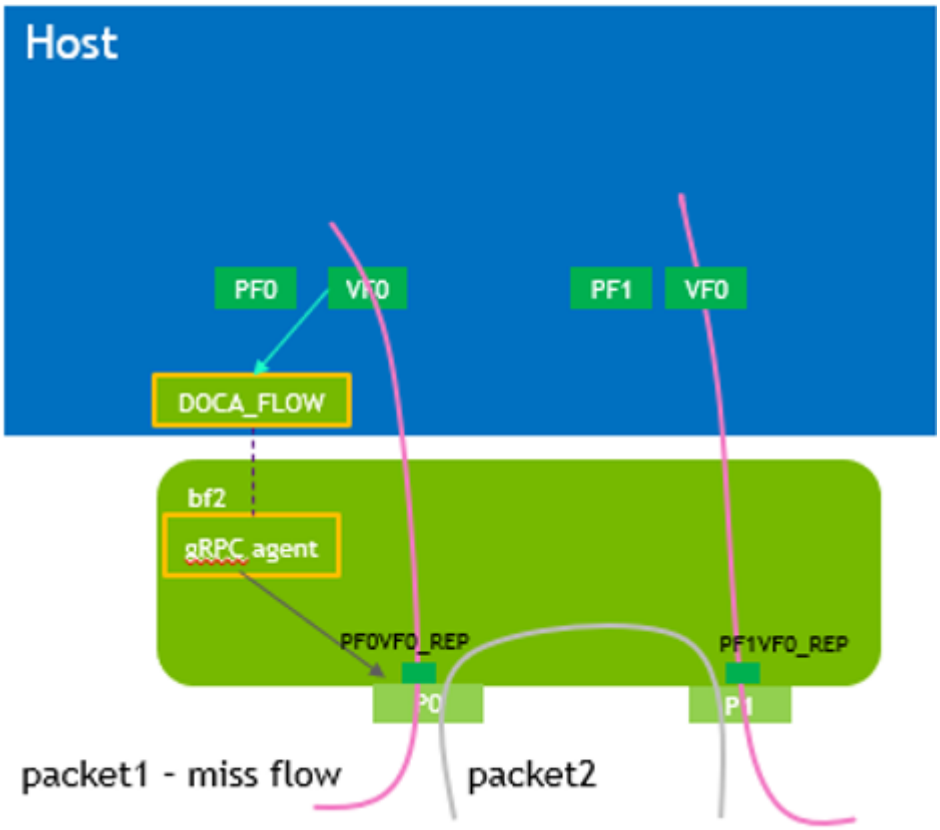
  The following diagram shows the basic flow of traffic in `switch` mode. Packet1 firstly misses to host RSS queues. The app captures this packet and decides which representor goes, and then sets the rule. Packets hit this rule and go to representor0.



- `remote-vnf`

  Remote mode is a BlueField mode only, with two physical ports (uplinks). Users must use `doca_flow_port_pair` to pair one physical port and one of its representors. A packet from this uplink, if it does not hit any rules from the users, is firstly received on this representor. Users must also use `doca_flow_port_pair` to pair two physical uplinks. If a packet is received from one uplink and hits the rule whose FWD action is to another uplink, then the packets are sent out from it.

  The following diagram shows the basic traffic flow in remote-vnf mode. Packet1, from BlueField uplink P0, firstly misses to host VF0. The app captures this packet and decides whether to drop it or forward it to another uplink (P1). Then, using gRPC to set rules on P0, packet2 hits the rule, then is either dropped or is sent out from P1.



5.2. Start Point

DOCA Flow API serves as an abstraction layer API for network acceleration. The packet processing in-network function is described from ingress to egress and, therefore, a pipe must be attached to the origin port. Once a packet arrives to the origin port, it starts the hardware execution as defined by the DOCA API.

`doca_flow_port` is an opaque object since the DOCA Flow API is not bound to a specific packet delivery API, such as DPDK. The first step is to start the DOCA Flow port by calling `doca_flow_port_start()`. The purpose of this step is to attach user application ports to the DOCA Flow ports.

When DPDK is used, the following configuration must be provided:

```
enum doca_flow_port_type type = DOCA_FLOW_PORT_DPDK_BY_ID;
const char *devargs = "1";
```

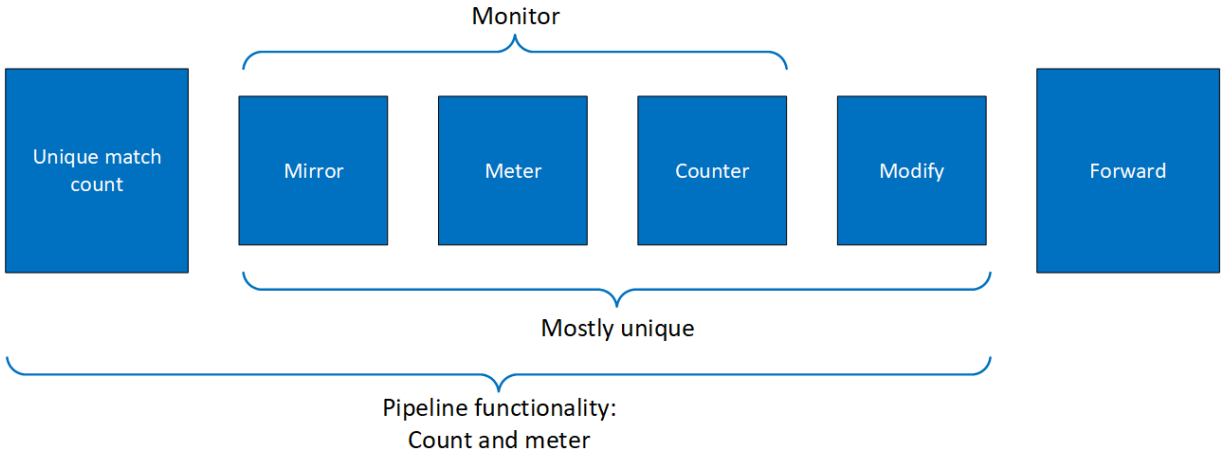
The `devargs` parameter points to a string that has the numeric value of the DPDK `port_id` in decimal format. The port must be configured and started before calling this API. Mapping the DPDK port to the DOCA port is required to synchronize application ports with hardware ports.

5.3. Create Pipe and Pipe Entry

Pipe is a template that defines packet processing without adding any specific HW rule. A pipe consists of a template that includes the following elements:

- Match
- Monitor
- Actions
- Forward

The following diagram illustrates a pipe structure.



The creation phase allows the HW to efficiently build the execution pipe. After the pipe is created, specific entries can be added. Only a subset of the pipe can be used (e.g. skipping the monitor completely, just using the counter, etc).

5.3.1. Setting Pipe Match

Match is a mandatory field when creating a pipe. Using the following struct, users must define the fields that should be matched on the pipe.

For each `doca_flow_match` field, users choose whether the field is:

- Ignored (wild card) – the value of the field is ignored
- Constant – all entries in the pipe must have the same value for this field. Users should not put a value for each entry.
- Changeable – per entry, the user must provide the value to match

> **Note:** L4 type, L3 type, and tunnel type cannot be changeable.

The match field type can be defined either implicitly or explicitly.

5.3.1.1. Implicit Match

To match implicitly, the following considerations should be taken into account.

- Ignored fields:
  - Field is zeroed
  - Pipeline has no comparison on the field
- Constant fields

  These are fields that have a constant value. For example, as shown in the following, the tunnel type is VXLAN.

  ```
  match.tun.type = DOCA_FLOW_TUN_VXLAN;
  ```

  These fields only need to be configured once, not once per new pipeline entry.
- Changeable fields

  These are fields that may change per entry. For example, the following shows an inner 5-tuple which are set with a full mask.

  ```
  match.in_dst_ip.ipv4_addr = 0xffffffff;
  ```

  If this is the constant value required by user, then they should set zero on the field when adding a new entry.
- Example

  The following is an example of a match on the VXLAN tunnel, where for each entry there is a specific IPv4 destination address, and an inner 5-tuple.

  ```
  static void build_underlay_overlay_match(struct doca_flow_match *match)
  {
      //outer
      match->out_dst_ip.ipv4_addr = 0xffffffff;
      match->out_l4_type = DOCA_PROTO_UDP;
      match->out_dst_port = DOCA_VXLAN_DEFAULT_PORT;
      match->tun.type = DOCA_FLOW_TUN_VXLAN;
      match->tun.vxlan_tun_id = 0xffffffff;
      //inner
      match->in_dst_ip.ipv4_addr = 0xffffffff;
      match->in_src_ip.ipv4_addr = 0xffffffff;
      match->in_src_ip.type = DOCA_FLOW_IP4_ADDR;
      match->in_l4_type = DOCA_PROTO_TCP;
      match->in_src_port = 0xffff;
      match->in_dst_port = 0xffff;
  }
  ```

5.3.1.2. Explicit Match

Users may provide a mask on a match. In this case, there are two `doca_flow_match` items: The first contains constant values and the second contains masks.

- Ignored fields
  - Field is zeroed
  - Pipeline has no comparison on the field

```
match_mask.in_dst_ip.ipv4_addr = 0;
```

- Constant fields

  These are fields that have a constant value. For example, as shown in the following, the tunnel type is VXLAN and the mask should be full.

  ```
  match.tun.type = DOCA_FLOW_TUN_VXLAN;
          match_mask.tun.type = 0xffffffff;
  ```

  Once a field is defined as constant, the field's value cannot be changed per entry. Users must set constant fields to zero when adding entries so as to avoid ambiguity.

- Changeable fields

  These are fields that may change per entry (e.g. inner 5-tuple). Their value should be zero and the mask should be full.

  ```
  match.in_dst_ip.ipv4_addr = 0;
          match_mask.in_dst_ip.ipv4_addr = 0xffffffff;
  ```

  Note that for IPs, the prefix mask can be used as well.

5.3.2. Setting Pipe Actions

5.3.2.1. Auto-modification

Similarly to setting pipe match, actions also have a template definition.

Similarly to `doca_flow_match` in the creation phase, only the subset of actions that should be executed per packet are defined. This is done in a similar way to match, namely by classifying a field of `doca_flow_match` to one of the following:

- Ignored field – field is zeroed, modify is not used
- Constant fields – when a field must be modified per packet, but the value is the same for all packets, a one-time value on action definitions can be used
- Changeable fields – fields that may have more than one possible value, and the exact values are set by the user per entry

  ```
  match_mask.in_dst_ip.ipv4_addr = 0xffffffff;
  ```

  Metadata is considered as per-packet changeable fields, pipe action is used as a mask.
- Boolean fields – Boolean values, encap and decap are considered as constant values. It is not allowed to generate actions with `encap=true` and to then have an entry without an encap value.

For example:

```
static void
create_decap_inner_modify_actions(struct doca_flow_actions *actions)
{
        actions->decap = true;
        actions->mod_dst_ip.ipv4_addr = 0xffffffff;
}
```

5.3.2.2. Explicit Modification Type

It is possible to force constant modification or per-entry modification with action description type (CONSTANT or SET) and mask. For example:

```
static void
create_constant_modify_actions(struct doca_flow_actions *actions#
```

```
                                struct doca_flow_action_descs *descs)
{
        actions->mod_src_port = 0x1234;
        descs->outer.src_port.type = DOCA_FLOW_ACTION_CONSTANT;
 descs->outer.src_port.mask.u64 = 0xffff;
}
```

5.3.2.3. Copy Field

Action description can be used to copy between packet field and metadata. For example:

```
static void
create_copy_packet_to_meta_actions(struct doca_flow_match *match#
                                   struct doca_flow_action_descs *descs)
{
        descs->outer.src_ip.type = DOCA_FLOW_ACTION_COPY;
        descs->outer.src_ip.copy.dst = &match->meta.u32[1];
}
```

5.3.2.4. Summary of Action Types

Pipe Creation			Entry Creation
action_desc		**Pipe Actions**	**Entry Actions**
doca_flow_action_type	**Configuration**		
`DOCA_FLOW_ACTION_AUTO` Derived from pipe actions.	No specific config	`0` – field ignored, no modification	N/A
		`val` != 0 – apply this val to all entries	N/A
		`val` = 0xfff – changeable field	Define `val` per entry
		Specific for Metadata - the meta field in the actions is used as a mask.	Define `val` per entry
`DOCA_FLOW_ACTION_CONSTANT` Pipe action is constant.	Define the mask	Define `val` to apply for all entries	N/A
`DOCA_FLOW_ACTION_SET` Set value from entry action.	Define the mask	N/A	Define `val` per entry
`DOCA_FLOW_ACTION_ADD` Add field value.	Define the `val` to apply for all entries	N/A	N/A
`DOCA_FLOW_ACTION_COPY` Copy field to another field.	Define the source and destination fields. ▶ Meta field -> header field ▶ Header field -> meta field ▶ Meta field -> meta field	N/A	N/A

If a meter policer should be used, then it is possible to have the same configuration for all policers on the pipe or to have a specific configuration per entry. The meter policer is determined by the FWD action. If an entry has NULL FWD action, the policer FWD action is taken from the pipe.

The monitor also includes the aging configuration, if the aging time is set, this entry ages out if timeout passes without any matching on the entry. User data is used to map user usage. If the `user_data` field is set, when the entry ages out, query API returns this `user_data`. If `user_data` is not configured by the application, the aged pipe entry handle is returned.

For example:

```
static void build_entry_monitor(struct doca_flow_monitor *monitor, void *user_ctx)
{
 monitor->flags |= DOCA_FLOW_MONITOR_AGING;
 monitor->aging = 10;
 monitor->user_data = (uint64_t)user_ctx;
}
```

Refer to Pipe Entry Aged Query for more information.

5.3.3. Setting Pipe Forwarding

The FWD (forwarding) "action" is the last action in a pipe, and it directs where the packet goes next. Users may configure one of the following destinations:

- Send to software (representor)
- Send to wire
- Jump to next pipe
- Drop packets

The FORWARDING action may be set for pipe create, but it can also be unique per entry.

A pipe can be defined with constant forwarding (e.g., always send packets on a specific port). In this case, all entries will have the exact same forwarding. If forwarding is not defined when a pipe is created, users must define forwarding per entry. In this instance, pipes may have different forwarding actions.

When a pipe includes meter monitor `<cir, cbs>`, it must have `fwd` defined as well as the policer.

The following is an RSS forwarding example:

```
fwd->type = DOCA_FLOW_FWD_RSS;
fwd->rss_queues = queues;
fwd->rss_flags = DOCA_FLOW_RSS_IP | DOCA_FLOW_RSS_UDP;
fwd->num_of_queues = 4;
fwd->rss_mark = 0x1234;
```

Queues point to the `uint16_t` array that contains the queue numbers. When a port is started, the number of queues is defined, starting from zero up to the number of queues minus 1. RSS queue numbers may contain any subset of those predefined queue numbers. For a specific match, a packet may be directed to a single queue by having RSS forwarding with a single queue.

MARK is an optional parameter that may be communicated to the software. If MARK is set and the packet arrives to the software, the value can be examined using the software API.

When DPDK is used, MARK is placed on the struct `rte_mbuf`. (See "Action: MARK" section in official DPDK documentation.) When using the Kernel, the MARK value is placed on the struct `sk_buff` MARK field.

The `port_id` is given in struct `doca_flow_port_cfg`.

The packet is directed to the port. In many instances the complete pipe is executed in the HW, including the forwarding of the packet back to the wire. The packet never arrives to the SW.

Example code for forwarding to port:

```
struct doca_flow_fwd *fwd = malloc(sizeof(struct doca_flow_fwd));
memset(fwd, 0, sizeof(struct doca_flow_fwd));
fwd->type = DOCA_FLOW_FWD_PORT;
fwd->port_id = port_cfg->port_id;
```

The type of forwarding is `DOCA_FLOW_FWD_PORT` and the only data required is the `port_id` as defined in `DOCA_FLOW_PORT`.

5.3.4. Pipe Create

Once all parameters are defined, the user should call `doca_flow_create_pipe` to create a pipe.

The return value of the function is a handle to the pipe. This handle should be given when adding entries to pipe. If a failure occurs, the function returns `NULL`, and the error reason and message are put in the error argument if provided by the user.

It is possible skip optional fields. For example, `fwd` and `fwd_miss` are not mandatory, and in pipe configuration some of the fields might be zeroed when not used. See Miss Pipe and Control Pipe for more information.

Once a pipe is created, a new entry can be added to it. These entries are bound to a pipe, so when a pipe is destroyed, all the entries in the pipe are removed. Please refer to section Pipe Entry for more information.

There is no priority between pipes or entries. The way that priority can be implemented is to match the highest priority first, and if a miss occurs, to jump to the next PIPE. There can be more than one PIPE on a root as long the pipes are not overlapping. If entries overlap, the priority is set according to the order of entries added. So, if two pipes have overlapping matching and PIPE1 has higher priority than PIPE2, users should add an entry to PIPE1 after any entry is added to PIPE2.

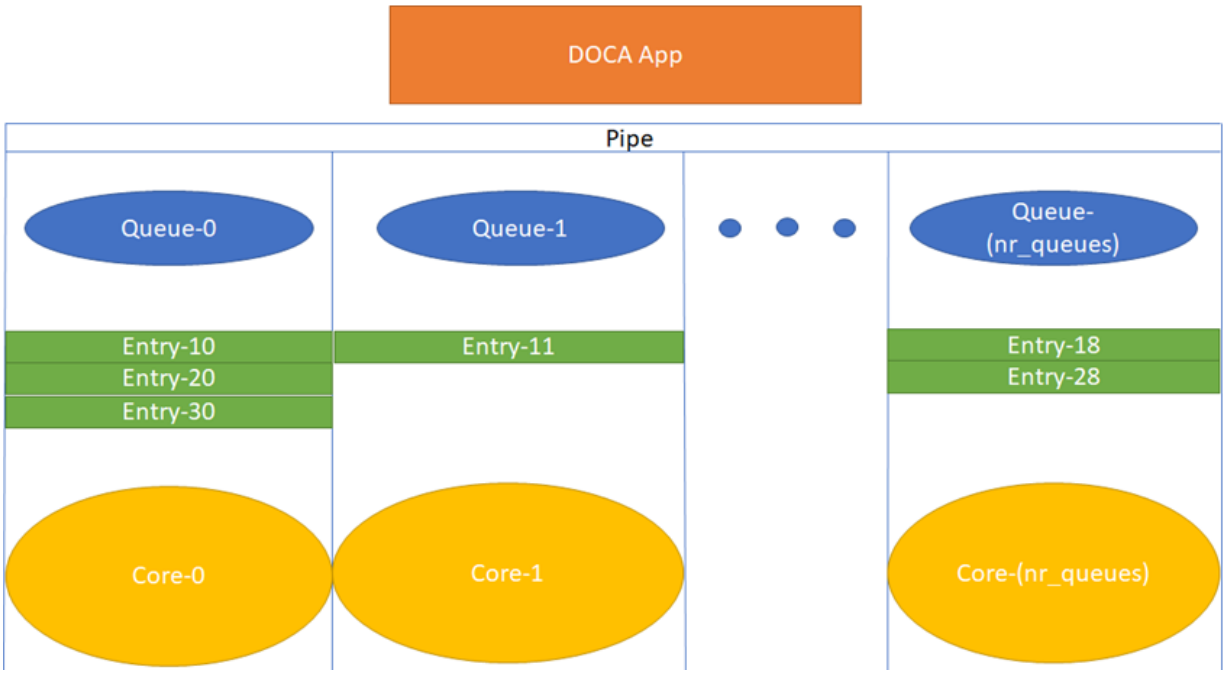
5.3.5. Pipe Entry (doca_flow_pipe_add_entry)

An entry is a specific instance inside of a pipe. When defining a pipe, users define match criteria (subset of fields to be matched), the type of actions to be done on matched packets, monitor, and, optionally, the FWD action.

When a user calls `doca_flow_pipe_add_entry()` to add an entry, they should define the values that are not constant among all entries in the pipe. And if FWD is not defined then that is also mandatory.

DOCA Flow is designed to support concurrency in an efficient way. Since the expected rate is going to be in millions of new entries per second, it is mandatory to use a similar architecture

as the data path. Having a unique queue ID per core saves the DOCA engine from having to lock the data structure and enables the usage of multiple queues when interacting with HW.



Each core is expected to use its own dedicated `pipe_queue` number when calling `doca_flow_pipe_entry`. Using the same `pipe_queue` from different cores causes a race condition and has unexpected results.

Upon success, a handle is returned. If a failure occurs, a NULL value is returned, and an error message is filled. The application can keep this handle and call remove on the entry using its handle.

```
int doca_flow_pipe_rm_entry(uint16_t pipe_queue, void *usr_ctx, struct
 doca_flow_pipe_entry *entry);
```

5.3.5.1. Pipe Entry Counting

By default, no counter is added. If defined in monitor, a unique counter is added per entry.

Note: Having a counter per entry affects performance and should be avoided if it is not required by the application.

When a counter is present, it is possible to query the flow and get the counter's data by calling `doca_flow_query`.

The retrieved statistics are stored in struct `doca_flow_query`.

5.3.5.2. Pipe Entry Aged Query

When a user calls `doca_flow_aged_query()`, this query is used to get the aged-out entries by the time quota in microseconds. The entry handle or the `user_data` input is returned by this API.

Since the number of flows can be very large, the query of aged flows is limited by a quota in microseconds. This means that it may return without all flows and requires the user to call it again. When the query has gone over all flows, a full cycle is done.

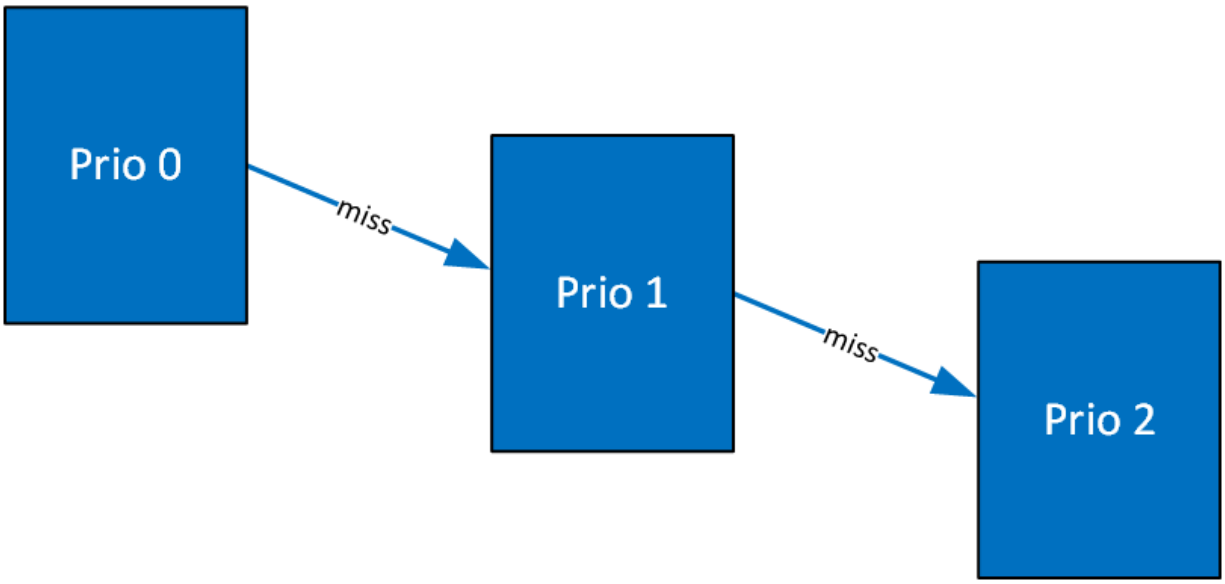
The struct `doca_flow_aged_query` contains the element `user_data` which contains the aged-out flow contexts.

5.3.6. Miss Pipe and Control Pipe

Note: On the root table, the user is not allowed to enter overlapping matches. If they do so, the match behavior is unpredictable. Multiple root tables are supported, but it is the user's responsibility to make sure entries do not overlap between root tables.

To set priority between pipes, users must use miss-pipes. Miss pipes allow to look up entries associated with pipe X, and if there are no matches, to jump to pipe X+1 and perform a lookup on entries associated with pipe X+1.

The following figure illustrates the HW table structure:



The first lookup is performed on the table with priority 0. If no hits are found, then it jumps to the next table and performs another lookup.

The way to implement miss-pipe in DOCA Flow is to use miss-pipe in FWD. In `struct doca_flow_fwd`, the field `miss_pipe` signifies that, when creating a pipe, if a miss-pipe

is configured then if a packet does not match the specific pipe, steering should jump to `miss_pipe`.

`miss_pipe` is defined as `doca_flow_pipe` and created by `doca_flow_create_pipe`. To separate `miss_pipe` and a general one, `is_root` is introduced in `struct doca_flow_pipe_cfg`. If `is_root` is true, it means the pipe is a root pipe executed on packet arrival. Otherwise, the pipe is `miss_pipe`.

When `fwd_miss` is not null, the packet that does not match the criteria is handled by `next_pipe` which is defined in `fwd_miss`.

In internal implementations of `doca_flow_create_pipe`, if `fwd_miss` is not null and the forwarding action type of `miss_pipe` is `DOCA_FLOW_FWD_PIPE`, a flow with the lowest priority is created that always jumps to the group for the `next_pipe` of the `fwd_miss`. Then the flow of `next_pipe` can handle the packets, or drop the packets if the forwarding action type of `miss_pipe` is `DOCA_FLOW_FWD_DROP`.

For example, VXLAN packets are forwarded as RSS and hairpin for other packets. The `miss_pipe` is for the other packets (non-VXLAN packets) and the match is for general Ethernet packets. The `fwd_miss` is defined by `miss_pipe` and the type is `DOCA_FLOW_FWD_PIPE`. For the VXLAN pipe, it is created by `doca_flow_create()` and `fwd_miss` is introduced.

Since, in the example, the jump flow is for general Ethernet packets, it is possible that some VXLAN packets match it and cause conflicts. For example, VXLAN flow entry for `ipA` is created. A VXLAN packet with `ipB` comes in, no flow entry is added for `ipB`, so it hits `miss_pipe` and is hairpinned.

A control pipe is introduced to handle the conflict. When a user calls `doca_flow_create_control_pipe()`, the new control pipe is created without any configuration except for the port. Then the user can add different matches with different forwarding and priorities when there are conflicts.

The user can add a control entry by calling `doca_flow_control_pipe_add_entry()`.

`priority` must be defined as higher than the lowest priority (3) and lower than the highest one (0).

The other parameters represent the same meaning of the parameters in `doca_flow_create_pipe`. In the example above, a control entry for VXLAN is created. The VLXAN packets with `ipB` hit the control entry.

5.3.7. Hardware Steering Mode

DPDK's Programmer's Guide describes the flow engine configuration for hardware steering mode. Users can enable hardware steering mode by following the instructions in DPDK's Network Interface Controller Drivers documentation.

Note: Follow this link for relevant release notes.

The following is an example of running DOCA with hardware steering mode:

```
.... -a 03:00.0, dv_flow_en=2 -a 03:00.1, dv_flow_en=2....
```

The following is an example of running DOCA with software steering mode:

```
.... -a 03:00.0 -a 03:00.1 ....
```

The `dv_flow_en=2` means that hardware steering mode is enabled.

In the struct `doca_flow_cfg`, the member `mode_args` represents DOCA applications. If it defined is with `hws` (e.g., `vnf,hws`, `switch,hws`, `remote_vnf,hws`) then hardware steering mode is enabled.

To create an entry by calling `doca_flow_pipe_add_entry`, the parameter `flags` can be set as `DOCA_FLOW_WAIT_FOR_BATCH` or `DOCA_FLOW_NO_WAIT`. `DOCA_FLOW_WAIT_FOR_BATCH` means that this flow entry waits to be pushed to hardware. Batch flows then can be pushed only at once. This reduces the push times and enhances the insertion rate. `DOCA_FLOW_NO_WAIT` means that the flow entry is pushed to hardware immediately.

The parameter `usr_ctx` is handled in the callback defined in struct `doca_flow_cfg`.

`doca_flow_entries_process` processes all the flows in this queue. After the flow is handled and the status is returned, the callback is executed with the status and `usr_ctx`.

If the user does not define the callback in `doca_flow_cfg`, the user can get the status using `doca_flow_entry_get_status` to check if the flow has completed offloading or not.

5.4. Teardown

5.4.1. Pipe Entry Teardown

When an entry is terminated by the user application or ages-out, the user should call the entry destroy function, `doca_flow_pipe_rm_entry()`. This frees the pipe entry and cancels hardware offload.

5.4.2. Pipe Teardown

Whena pipe is terminated by the user application, the user should call the pipe destroy function, `doca_flow_destroy_pipe()`. This destroys the pipe and the pipe entries that match it.

When all pipes of a port are terminated by the user application, the user should call the pipe flush function, `doca_flow_port_pipe_flush()`. This destroys all pipes and all pipe entries belonging to this port.

5.4.3. Port Teardown

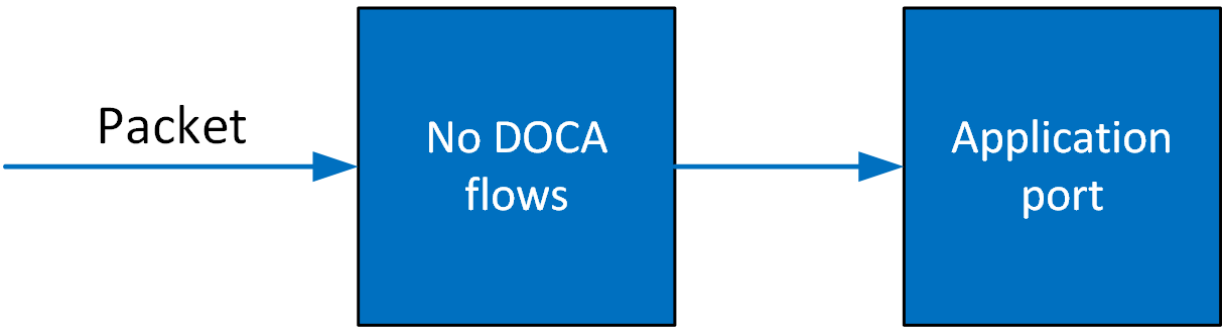
When the port is not used anymore, the user should call the port destroy function, `doca_flow_destroy_port()`. This destroys the DOCA port and frees all resources of the port.

5.4.4. Flow Teardown

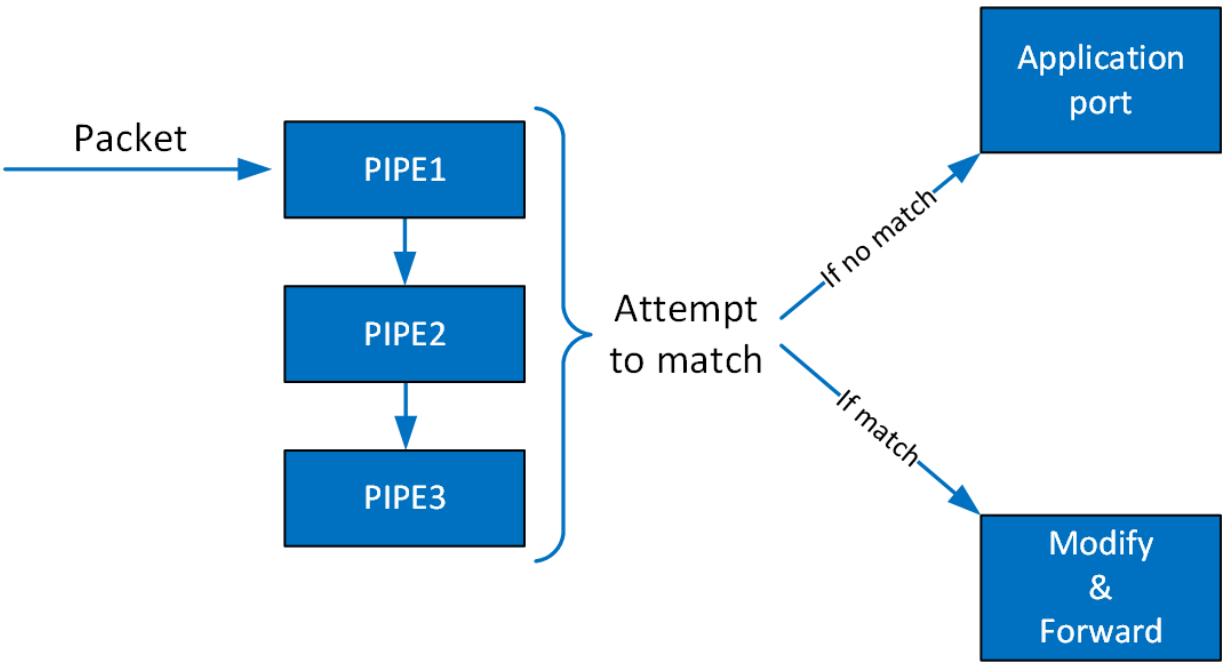
When the DOCA Flow is not used anymore, the user should call the flow destroy function, `doca_flow_destroy()`. This releases all the resources used by DOCA Flow.

Chapter 6. Packet Processing

In situations where there is a port without a pipe defined, or with a pipe defined but without any entry, the default behavior is that all packets arrive to a port in the software.



Once entries are added to the pipe, if a packet has no match then it continues to the port in the software. If it is matched, then the rules defined in the pipe are executed.



If the packet is forwarded in RSS, the packet is forwarded to software according to the RSS definition. If the packet is forwarded to a port, the packet is redirected back to the wire. If the packet is forwarded to the next pipe, then the software attempts to match it with the next pipe.

Note that the number of pipes impacts performance. The longer the number of matches and actions that the packet goes through, the longer it takes the HW to process it. When there is a very large number of entries, the HW needs to access the main memory to retrieve the entry context which increases latency.

Chapter 7. DOCA Flow gRPC

This chapter describes gRPC support for DOCA Flow. The DOCA Flow gRPC-based API allows users on the host to leverage the HW offload capabilities of the BlueField DPU using gRPCs from the host itself.

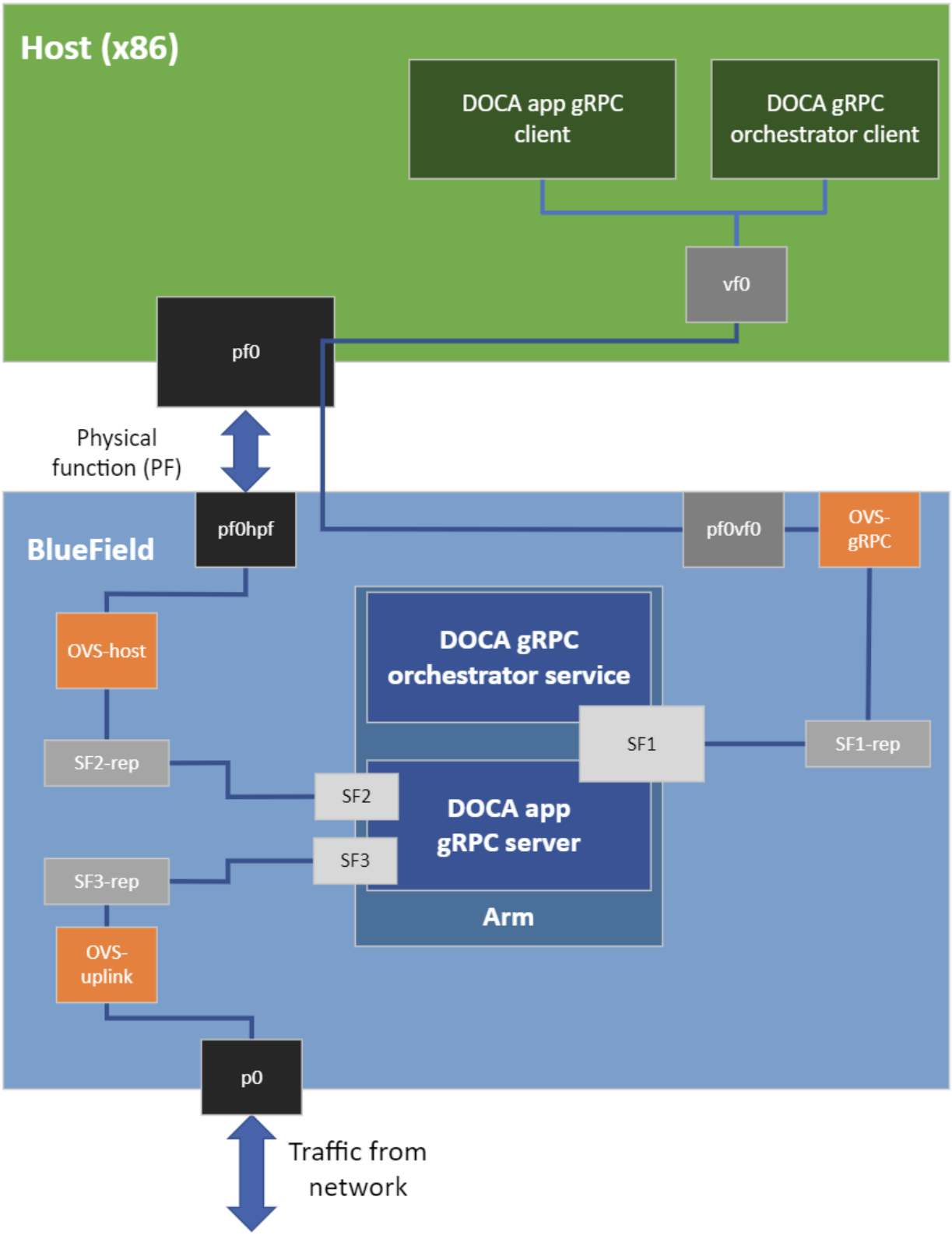
DOCA Flow gRPC server implementation is based on gRPC's `async` API to maximize the performance offered to the gRPC client on the host. In addition, the gRPC support in the DOCA Flow library provides a client interface which gives the user the ability to send/receive messages to/from the client application in C.

This section is divided into the following parts:

- proto-buff – this section details the messages defined in the proto-buff
- Client interface – this section details the API for communicating with the server
- Usage – this section explains how to use the client interface to develop your own client application based on DOCA Flow gRPC support

Refer to NVIDIA DOCA gRPC Infrastructure User Guide for more information about DOCA gRPC support.

The following figure illustrates the DOCA Flow gRPC server-client communication when running in VNF mode.


Host (x86)
DOCA app gRPC client
DOCA gRPC orchestrator client
vf0
pf0
Physical function (PF)
BlueField
pf0hpf
pf0vf0
OVS-gRPC
OVS-host
DOCA gRPC orchestrator service
SF1
SF1-rep
SF2-rep
SF2
DOCA app gRPC server
SF3
SF3-rep
Arm
OVS-uplink
p0
Traffic from network

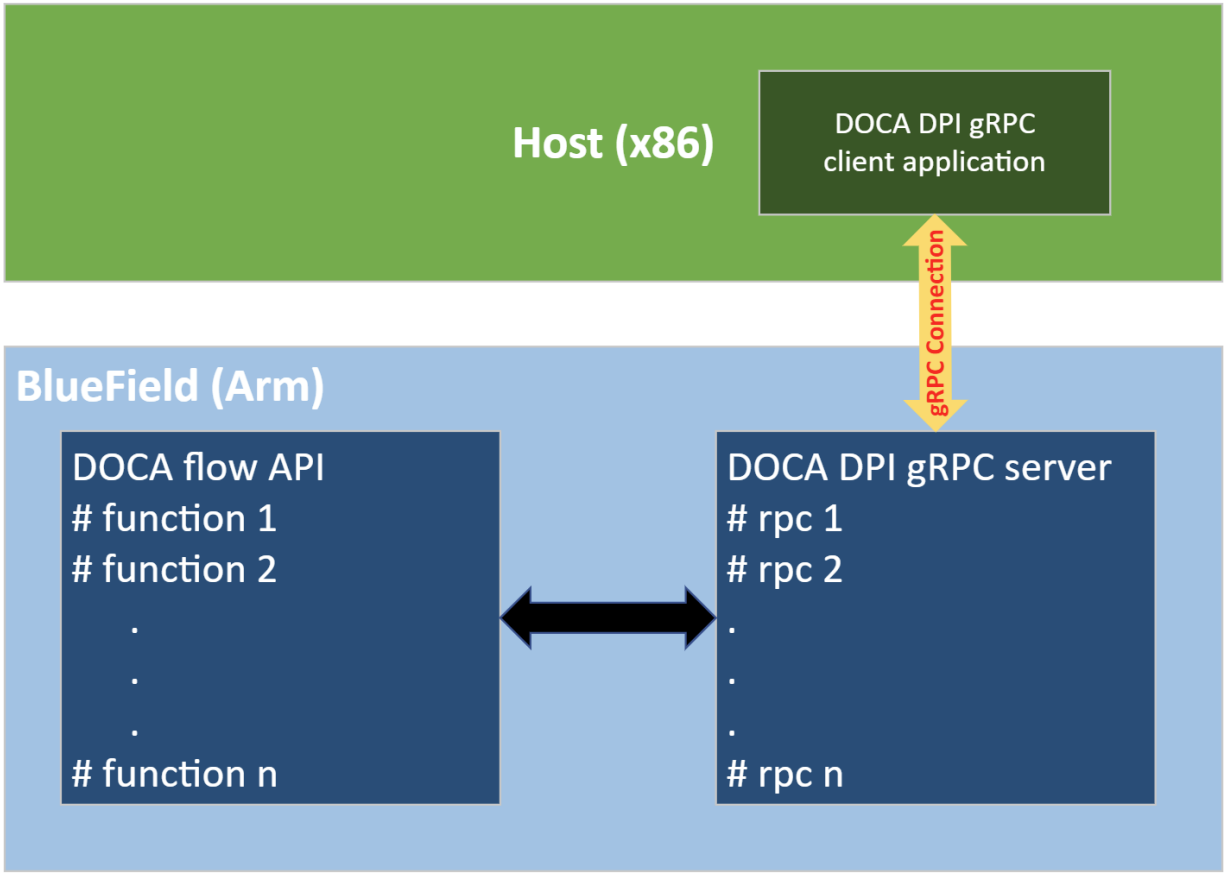
7.1. Proto-Buff

As with every gRPC proto-buff, DOCA Flow gRPC proto-buff defines the services it introduces, and the messages used for the communication between the client and the server. Each proto-buff DOCA Flow method:

- Represents exactly one function in DOCA Flow API
- Has its request message, depending on the type of the service
- Has the same response message (`DocaFlowResponse`)

In addition, DOCA Flow gRPC proto-buff defines several of messages that are used for defining request messages, the response message, or other messages.

Each message defined in the proto-buff represents either a struct or an enum defined by DOCA Flow API. The following figure illustrates how DOCA Flow gRPC server represents the DOCA Flow API.



The proto-buff path for DOCA Flow gRPC is `/opt/mellanox/doca/infrastructure/doca_grpc/doca_flow/doca_flow.proto`.

7.1.1. Response Message

All services have the same response message. `DocaFlowResponse` contains all types of results that the services may return to the client.

```
/** General DOCA Flow response message */
message DocaFlowResponse{
    bool success = 1;    /* True in case of success */
    DocaFlowError error = 2; /* Otherwise, this field contains the error information
 */
    /* in case of success, one or more of the following may be used */
        uint32 port_id = 3;
        uint64 pipe_id = 4;
        uint64 entry_id = 5;
        string port_pipes_dump = 6;
        DocaFlowQueryRes query_res = 7;
        bytes priv_data = 8;
        DocaFlowHandleAgingRes handle_aging_res = 9;
 uint64 nb_entries_processed = 10;
 DocaFlowEntryStatus status = 11;
}
```

7.1.2. DocaFlowCfg

The `DocaFlowCfg` message represents the `doca_flow_cfg` struct.

7.1.3. DocaFlowPortCfg

The `DocaFlowPortCfg` message represents the `doca_flow_port_cfg` struct.

7.1.4. DocaFlowPipeCfg

The `DocaFlowPipeCfg` message represents the `doca_flow_pipe_cfg` struct.

7.1.5. DocaFlowMatch

The `DocaFlowMatch` message represents the `doca_flow_match` struct.

The `DocaFlowMatch` message contains fields of types `DocaFlowIPAddress` and `DocaFlowTun`. These types are messages which are also defined in the `doca_flow.proto` file and represents `doca_flow_ip_address` and `doca_flow_tun` respectively.

7.1.6. DocaFlowActions

The `DocaFlowActions` message represents the `doca_flow_actions` struct.

Like the `DocaFlowMatch` message, the `DocaFlowActions` message also contains fields of type `DocaFlowIPAddress` to represent modify actions on a source or destination IP addresses.

7.1.7. DocaFlowMonitor

The `DocaFlowMonitor` message represents the `doca_flow_monitor` struct.

7.1.8. DocaFlowQueryStats

The `DocaFlowQueryStats` message represents the `doca_flow_query` struct.

7.1.9. EnvInit

Before using any DOCA Flow method, it is mandatory to call environment initialization gRPC to initialize the necessary resources used by the underlying API that DOCA Flow uses for offloading the hardware rules. For example, if DPDK is the underlying API that is used, then calling the `EnvInit` gRPC initializes all required DPDK resources (ports, queues, etc).

Environment initialization gRPC:

```
rpc EnvInit(EnvCfg) returns (DocaFlowResponse);
```

`EnvCfg` contains the needed details for initializing the environment.

If successful, the `success` field in the response message is set to `true`. Otherwise, the `error` field is populated with the error information.

7.1.10. DocaFlowInit

DOCA Flow initialization gRPC:

```
rpc DocaFlowInit(DocaFlowCfg) returns (DocaFlowResponse);
```

If successful, the `success` field in the response message is set to `true`. Otherwise, the `error` field is populated with the error information.

7.1.11. DocaFlowPortStart

The service for starting the DOCA flow ports:

```
rpc DocaFlowPortStart(DocaFlowPortCfg) returns (DocaFlowResponse);
```

If successful, the `success` field in the `DocaFlowResponse` is set to `true`. Otherwise, the `error` field is populated with the error information.

7.1.12. DocaFlowCreatePipe

The `DocaFlowCreatePipeRequest` message contains all the necessary information for pipe creation as the DOCA Flow API suggests:

```
message DocaFlowCreatePipeRequest {
        DocaFlowPipeCfg cfg = 1;            /* the pipe configurations */
        DocaFlowFwd fwd = 2;                /* the pipe's FORWARDING component  */
        DocaFlowFwd fwd_miss = 3;           /* The FORWARDING miss component */
}
```

Once all the parameters are defined, a "create pipe" service can be called:

```
rpc DocaFlowCreatePipe (DocaFlowCreatePipeRequest) returns (DocaFlowResponse);
```

If successful, the `success` field in `DocaFlowResponse` is set to `true`, and the `pipe_id` field is populated with the ID of the created pipe. This ID should be given when adding entries to the pipe. Otherwise, the error field is filled accordingly.

7.1.13. DocaFlowPipeAddEntry

The `DocaFlowPipeAddEntryRequest` message contains all the necessary information for adding an entry to the pipe:

```
message DocaFlowPipeAddEntryRequest{
        uint32 port_id = 1;                 /* the port ID to add the entry to */
        uint32 pipe_queue = 2;              /* the pipe queue */
        uint64 pipe_id = 3;                 /* the pipe ID to add the entry to */
        DocaFlowMatch match = 4;            /* matcher for the entry */
        DocaFlowActions actions = 5;        /* actions for the entry */
        DocaFlowMonitor monitor = 6;        /* monitor for the entry */
        DocaFlowFwd fwd = 7;                /* The entry's FORWARDING component */
}
```

Once all the parameters are defined, an "add entry to pipe" service can be called:

```
rpc DocaFlowPipeAddEntry(DocaFlowPipeAddEntryRequest) returns (DocaFlowResponse);
```

If successful, the `success` field in `DocaFlowResponse` is set to `true`, and the `entry_id` field is populated with the ID of the added entry. This ID should be given when adding entries to the pipe. Otherwise, the error field is filled accordingly.

7.1.14. DocaFlowControlPipeAddEntry

The `DocaFlowControlPipeAddEntryRequest` message contains the required arguments for adding entries to the control pipe:

```
message DocaFlowControlPipeAddEntryRequest{
        uint32 port_id = 1;                 /* the port ID to add the entry to */
        uint32 priority = 2;                /* he priority of the added entry to the
 filter pipe */
        uint32 pipe_queue = 3;              /* the pipe queue */
        uint32 pipe_id = 4;                 /* the pipe ID to add the entry to */
        DocaFlowMatch match = 5;            /* matcher for the entry */
        DocaFlowMatch match_mask = 6;       /* matcher mask for the entry */
        DocaFlowFwd fwd = 7;                /* The entry's FORWARDING component */
}
```

Once all the parameters are defined, an "add entry to pipe" service can be called:

```
rpc DocaFlowControlPipeAddEntry(DocaFlowControlPipeAddEntryRequest) returns
 (DocaFlowResponse);
```

If successful, the `success` field in `DocaFlowResponse` is set to `true`, and the `entry_id` field is populated with the ID of the added entry. This ID should be given when adding entries to the pipe. Otherwise, the error field is filled accordingly.

7.1.15. DocaFlowEntriesProcess

The `DocaFlowEntriesProcessRequest` contains the required arguments for processing the entries in the queue.

```
message DocaFlowEntriesProcessRequest{
        uint32 port_id = 1;          /* the port ID of the entries to process. */
        uint32 pipe_queue = 2;       /* the pipe queue of the entries to process.
 */
        /* max time in micro seconds for the actual API to process entries. */
        uint64 timeout = 3;
 /* An upper bound for the required number of entries to process. */
        uint32 max_processed_entries = 4;
}
```

Once all the parameters are defined, the "entries process" service can be called:

```
rpc DocaFlowEntriesProcess(DocaFlowEntriesProcessRequest) returns
 (DocaFlowResponse);
```

If successful, the `success` field in `DocaFlowResponse` is set to `true`, and the `nb_entries_processed` field is populated with the ID of the number of processed entries.

7.1.16. DocaFlowEntyGetStatus

The `DocaFlowEntryGetStatusRequest` contains the required arguments for fetching the status of a given entry.

```
message DocaFlowEntryGetStatusRequest{
        /* the entry identifier of the requested entry's status. */
        uint64 entry_id = 1;
}
```

Once all the parameters are defined, the "entry get status" service can be called:

```
rpc DocaFlowEntriesProcess(DocaFlowEntriesProcessRequest) returns
 (DocaFlowResponse);
```

If successful, the `success` field in `DocaFlowResponse` is set to `true`, and the `status` field is populated with the status of the requested entry. This field's type is `DocaFlowEntryStatus`, which is an enum defined in the proto-buff, and represents the enum `doca_flow_entry_status`, defined in the DOCA Flow header.

7.2. DOCA Flow gRPC Client API

This section describes the recommended way for C developers to utilize gRPC support for DOCA Flow API. Refer to the DOCA Flow gRPC API in NVIDIA DOCA Libraries API Reference Manual for the library API reference.

The following sections provide additional details about the library API.

The DOCA installation includes `libdoca_flow_grpc` which is a library that provides a C API wrapper to the C++ gRPC, while mimicking the regular DOCA Flow API, for ease of use, and allowing smooth transition to the Arm.

This library API is exposed in `doca_flow_grpc_client.h` and is essentially the same as `doca_flow.h`, with the notation differences detailed in the following subsections. In general, the client interface API usage is almost identical to the regular API (i.e., DOCA Flow API). The arguments of each function in DOCA Flow API, are almost identical to the arguments of each function defined in the client API, except that each pointer is replaced with an ID representing the pointer.

For example, when creating a pipe or adding an entry, the original API returns a pointer to the created pipe or the added entry. However, when adding an entry or creating a pipe using the client interface, an ID representing the added entry or the created pipe is returned to the client application instead of the pointer.

7.2.1. doca_flow_grpc_response

`doca_flow_grpc_response` is a general response struct that holds information regarding the function result. Each API returns this struct. If an error occurs, the `error` field is populated

with the error's information, and the `success` field is set to `false`. Otherwise, the `success` field is set to `true` and one of the other fields may hold a return value depending on the called function.

For example, when calling `doca_flow_grpc_create_pipe()` the `pipe_id` field is populated with the ID of the created pipe in case of success.

```
struct doca_flow_grpc_response {
      bool success;
      struct doca_flow_error error;
      uint64_t pipe_id;
      uint64_t entry_id;
      uint32_t aging_res;
      uint64_t nb_entries_processed;
      enum doca_flow_entry_status entry_status;
      };
```

success
In case of success, the value should be `true`.

error
In case of error, this struct should contain the error information.

pipe_id
Pipe ID of the created pipe.

entry_id
Entry ID of the created entry.

aging_res
Return value from handle aging.

nb_entries_processed
Return value from entries process.

entry_status
Return value from entry get status.

7.2.2. doca_flow_grpc_pipe_cfg

`doca_flow_grpc_pipe_cfg` is a pipeline configuration wrapper.

```
struct doca_flow_grpc_pipe_cfg {
 struct doca_flow_pipe_cfg cfg;
 uint16_t port_id;
};
```

cfg
Pipe configuration containing the user-defined template for the packet process.

port_id
Port ID for the pipeline.

7.2.3. doca_flow_grpc_fwd

`doca_flow_grpc_fwd` is a forwarding configuration wrapper.

```
struct doca_flow_grpc_fwd {
 struct doca_flow_fwd fwd;
 uint64_t next_pipe_id;
};
```

fwd
Forward configuration which directs where the packet goes next.

next_pipe_id
When using `DOCA_FLOW_FWD_PIPE`, this field contains the next pipe's ID.

7.2.4. doca_flow_grpc_client_create

This function initializes a channel to DOCA Flow gRPC server.

This must be invoked first before any other function in this API. This is a one-time call.

```
void doca_flow_grpc_client_create(char *grpc_address);
```

grpc_address [in]
String representing the server IP.

7.2.5. doca_flow_grpc_env_init

This function initializes ports and queues, allocates mempool, and binds hairpin queues of each port to its peer port.

Users must first initialize the underlying resources, which is used by the underlying API used by DOCA Flow, to add flow rules.

This function must be invoked after creating the client and before any other DOCA Flow client API.

```
struct doca_flow_grpc_response doca_flow_grpc_env_init(struct
 application_dpdk_config *dpdk_config);
```

dpdk_config [in]
Pointer to `application_dpdk_config`, holding some initialization configuration information.

Returns
`doca_flow_grpc_response` struct.

7.2.6. doca_flow_grpc_env_destroy

This function destroys the underlying environment resources.

```
struct doca_flow_grpc_response doca_flow_grpc_env_init(void);
```

Returns
`doca_flow_grpc_response` struct.

7.3. DOCA Flow gRPC Usage

A DOCA flow gRPC based server is implemented using the `async` API of gRPC. This is because the `async` API gives the server the ability to expose DOCA flow's concurrency support. Therefore, it is very important to use the client interface API for communicating with the DOCA Flow gRPC server because it hides all gRPC-related details from the users, which eases the use of the server, and exposes to the client applications the efficiency of DOCA Flow, in terms of flow insertion rates.

The following phases demonstrate a basic flow of client applications:

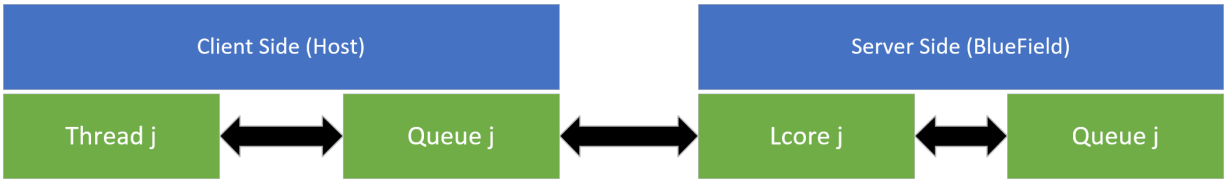
- Init Phase – client interface and environment initializations

- Flow life cycle – this phase is the same phase described in chapter Flow Life Cycle

It is important to emphasize that the number of threads for adding entries should be the same as the number of queues used when starting the server and initializing the environment (DPDK) and DOCA Flow API. This is to prevent bottlenecks on the server side.

If a client application starts the server on BlueField with N cores (through EAL arguments), this means that environment and DOCA Flow initialization should be done with N queues. As a result, the server launches N lcores, each one responsible for exactly one queue that is accessed only by it. Therefore, the client application should launch N threads as well, each being responsible for adding entries to a specific queue which is accessed by it only as well.

The following illustration demonstrates the relation between thread "j" on the client side and lcore "j" on the server side:



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