RDMA support in SONiC

Wenda Ni

Lossless fabric

- PFC frame generation
- PFC storm watchdog
- DCQCN support

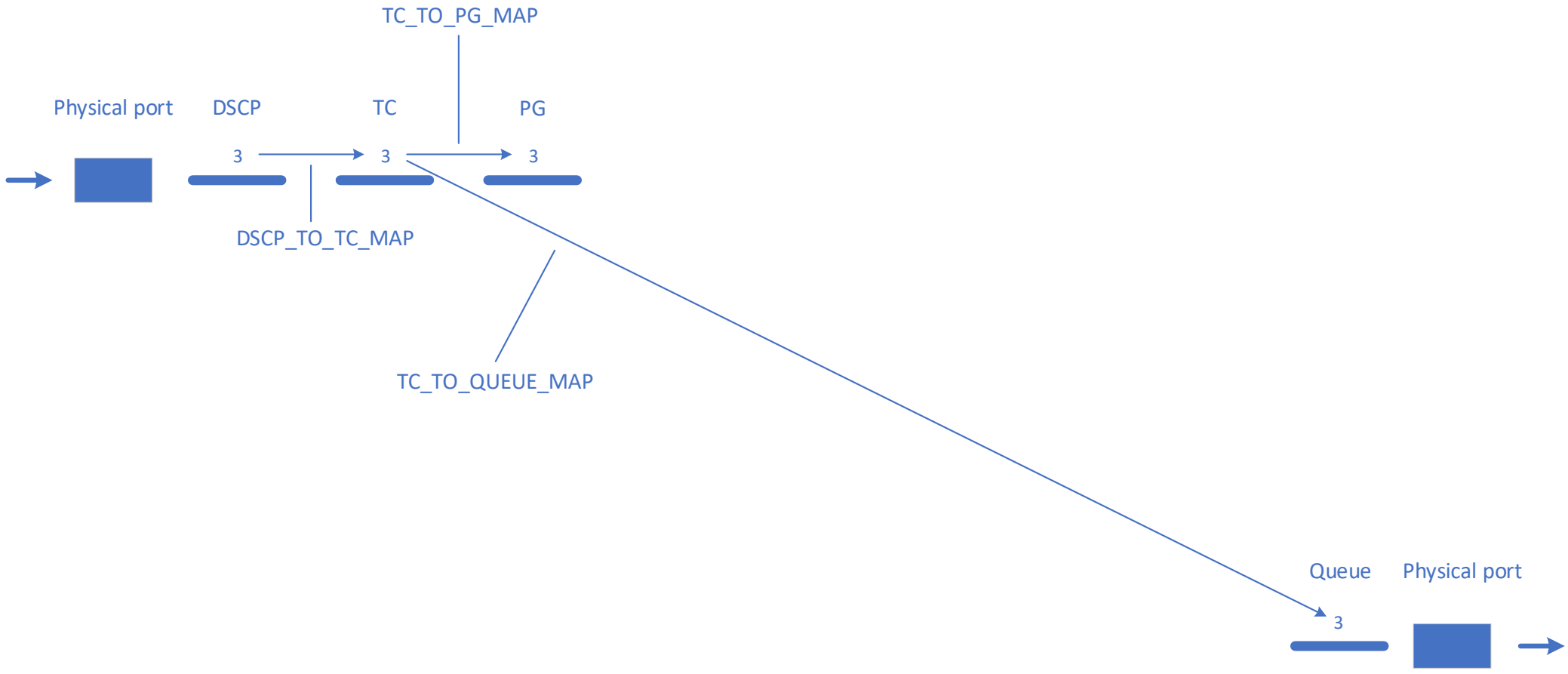
PFC frame generation

- QoS mapping
- Memory management unit (MMU)
- PFC generation
- PFC reception

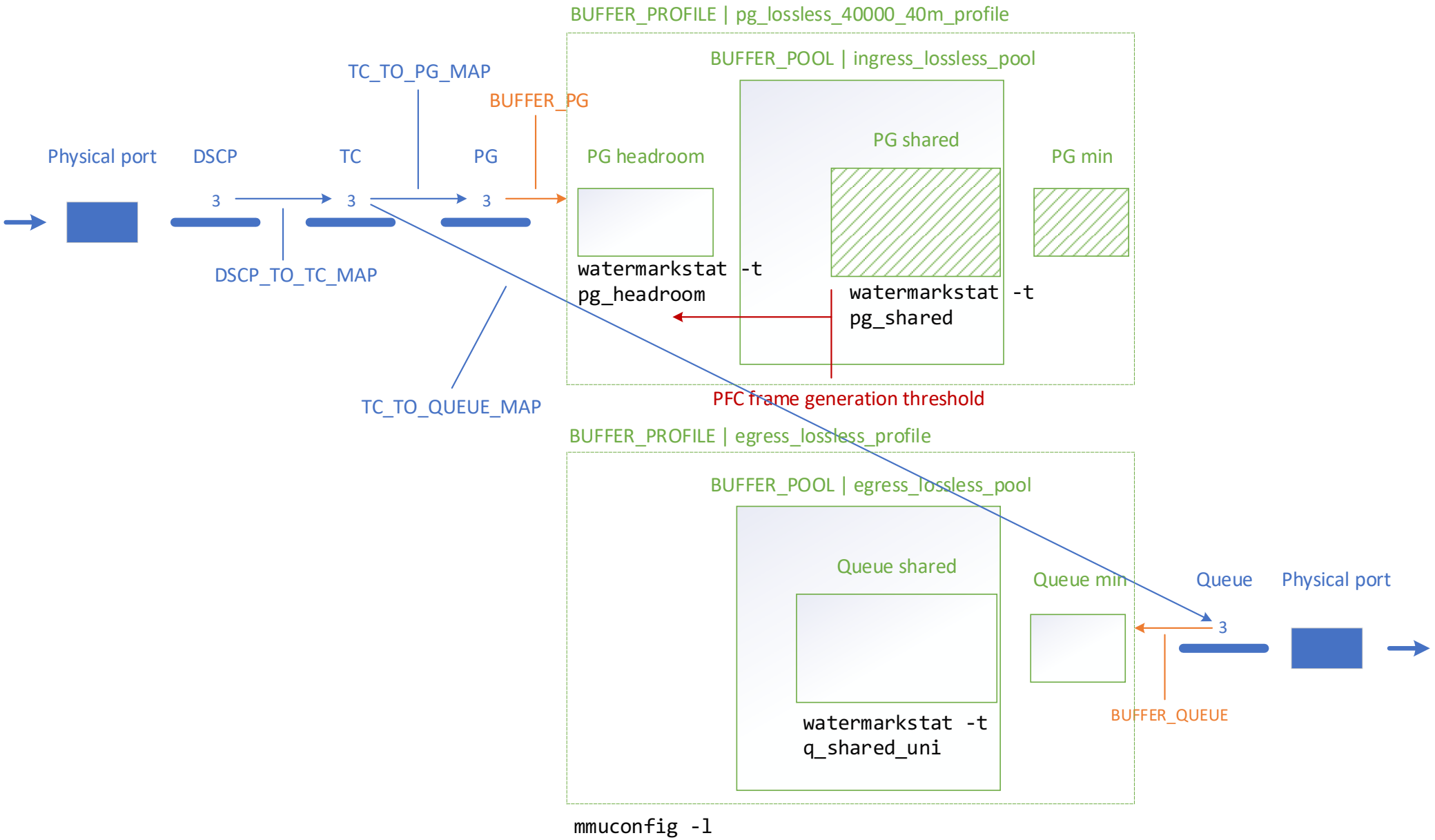
PFC frames

```
"""
Set PFC defined fields and generate the packet

The Ethernet Frame format for PFC packets is the following:

Destination MAC |   01:80:C2:00:00:01   |
                ------------------------
Source MAC      |      Station MAC      |
                ------------------------
Ethertype       |         0x8808        |
                ------------------------
OpCode          |         0x0101        |
                ------------------------
Class Enable V  | 0x00 E7...E0          |
                ------------------------
Time Class 0    |       0x0000          |
                ------------------------
Time Class 1    |       0x0000          |
                ------------------------
...
                ------------------------
Time Class 7    |       0x0000          |
                ------------------------
"""
```

QoS mapping: ingress PG & egress queue



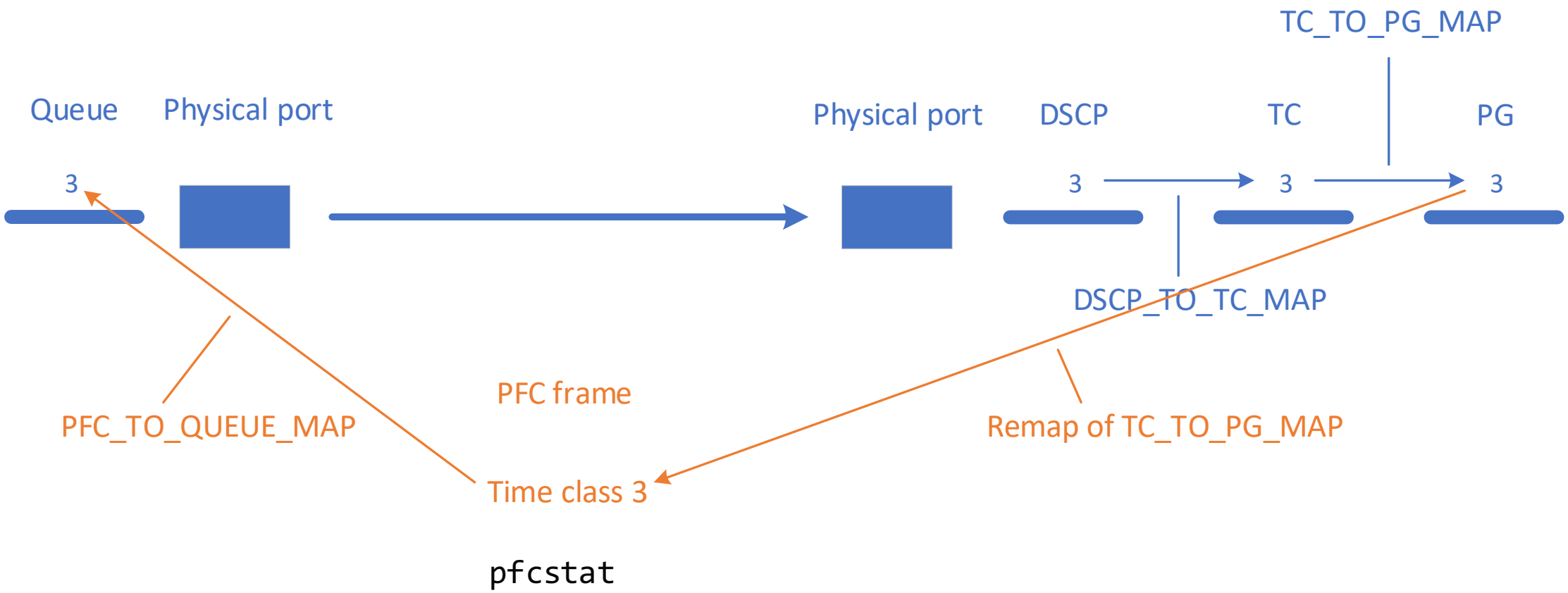
Attach buffers to ingress pg & egress queue



PFC frame stats

```
admin@str-a7050-acs-1:~$ pfcstat
    Port Rx    PFC0    PFC1    PFC2     PFC3    PFC4    PFC5    PFC6    PFC7
-----------  ------  ------  ------  -------  ------  ------  ------  ------
  Ethernet0       0       0       0        0       0       0       0       0
  Ethernet4       0       0       0        0       0       0       0       0
  Ethernet8       0       0       0        0       0       0       0       0
 Ethernet12       0       0       0        0       0       0       0       0
 Ethernet16       0       0       0        0       0       0       0       0
 Ethernet20       0       0       0        0       0       0       0       0
 Ethernet24       0       0       0        0       0       0       0       0
 Ethernet28       0       0       0  2124135       0       0       0       0
 Ethernet32       0       0       0        0       0       0       0       0
 Ethernet36       0       0       0        0       0       0       0       0
 Ethernet40       0       0       0        0       0       0       0       0
 Ethernet44       0       0       0        0       0       0       0       0
 Ethernet48       0       0       0        0       0       0       0       0
 Ethernet52       0       0       0        0       0       0       0       0
 Ethernet56       0       0       0        0       0       0       0       0
 Ethernet60       0       0       0        0       0       0       0       0
 Ethernet64       0       0       0        0       0       0       0       0
 Ethernet68       0       0       0        0       0       0       0       0
 Ethernet72       0       0       0        0       0       0       0       0
```

PFC reception



`pfcstat`

PFC storm watchdog

- Confirm a queue is in storm
- Detection & restoration logic

Confirm a queue is in storm

- Approach 1 (indirect)
  - Tx & Rx drop increasing

```
admin@str-a7050-acs-1:~$ pfcwd show stats
       QUEUE    STORM DETECTED/RESTORED    TX OK/DROP    RX OK/DROP    TX LAST OK/DROP    RX LAST OK/DROP
------------  -------------------------  ------------  ------------  -----------------  -----------------
Ethernet28:3                         2/1           0/0           0/0                0/0                0/0
```

Confirm a queue is in storm

- Approach 2 (direct)
  - Step 1: confirm the orchagent process is alive

```
admin@str-a7050-acs-1:~$ ps aux | grep orch
root      5399  0.1  0.6 201320 25480 pts/0    Sl   Mar28   0:18 /usr/bin/orchagent -d /var/log/swss -b 8192 -m 28:99:3a:20:8e:48
```

Confirm a queue is in storm

- Approach 2
  - Step 1: confirm the orchagent process is alive
  - Step 2: Obtain the oid of the queue

```
admin@str-a7050-acs-1:~$ redis-cli -n 2 hget "COUNTERS_QUEUE_NAME_MAP" "Ethernet28:3"
"oid:0x150000000001bf"
```

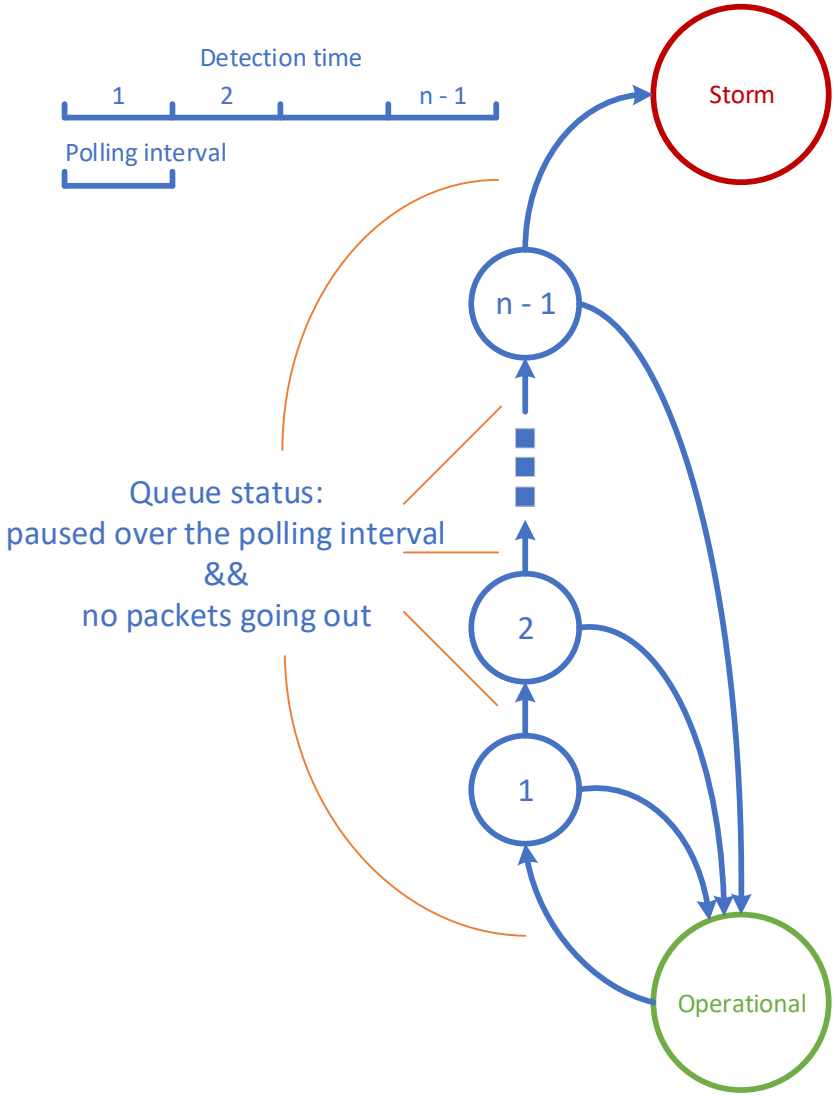
Confirm a queue is in storm

- Approach 2
  - Step 1: confirm the orchagent process is alive
  - Step 2: Obtain the oid of the queue
  - Step 3: Query the counters
    - Line 11) & 12)

```
admin@str-a7050-acs-1:~$ redis-cli -n 2 hgetall "COUNTERS:oid:0x150000000001bf"
 1) "PFC_WD_DETECTION_TIME"
 2) "200000"
 3) "PFC_WD_RESTORATION_TIME"
 4) "200000"
 5) "PFC_WD_ACTION"
 6) "drop"
 7) "PFC_WD_QUEUE_STATS_DEADLOCK_DETECTED"
 8) "2"
 9) "PFC_WD_QUEUE_STATS_DEADLOCK_RESTORED"
10) "1"
11) "PFC_WD_STATUS"
12) "stormed"
13) "SAI_QUEUE_STAT_PACKETS"
14) "0"
15) "SAI_QUEUE_STAT_CURR_OCCUPANCY_BYTES"
16) "0"
17) "SAI_QUEUE_ATTR_PAUSE_STATUS"
18) "true"
19) "SAI_QUEUE_ATTR_PAUSE_STATUS_last"
20) "true"
21) "SAI_QUEUE_STAT_PACKETS_last"
22) "0"
23) "PFC_WD_DETECTION_TIME_LEFT"
24) "200000"
25) "PFC_WD_QUEUE_STATS_TX_PACKETS"
26) "0"
27) "PFC_WD_QUEUE_STATS_TX_DROPPED_PACKETS"
28) "0"
29) "PFC_WD_QUEUE_STATS_RX_PACKETS"
30) "0"
31) "PFC_WD_QUEUE_STATS_RX_DROPPED_PACKETS"
32) "0"
33) "PFC_WD_QUEUE_STATS_TX_PACKETS_LAST"
34) "0"
35) "PFC_WD_QUEUE_STATS_TX_DROPPED_PACKETS_LAST"
36) "0"
37) "PFC_WD_QUEUE_STATS_RX_PACKETS_LAST"
38) "0"
39) "PFC_WD_QUEUE_STATS_RX_DROPPED_PACKETS_LAST"
40) "0"
41) "PFC_WD_RESTORATION_TIME_LEFT"
42) "200000"
43) "SAI_QUEUE_STAT_BYTES"
44) "0"
45) "SAI_QUEUE_STAT_DROPPED_PACKETS"
46) "0"
47) "SAI_QUEUE_STAT_DROPPED_BYTES"
```

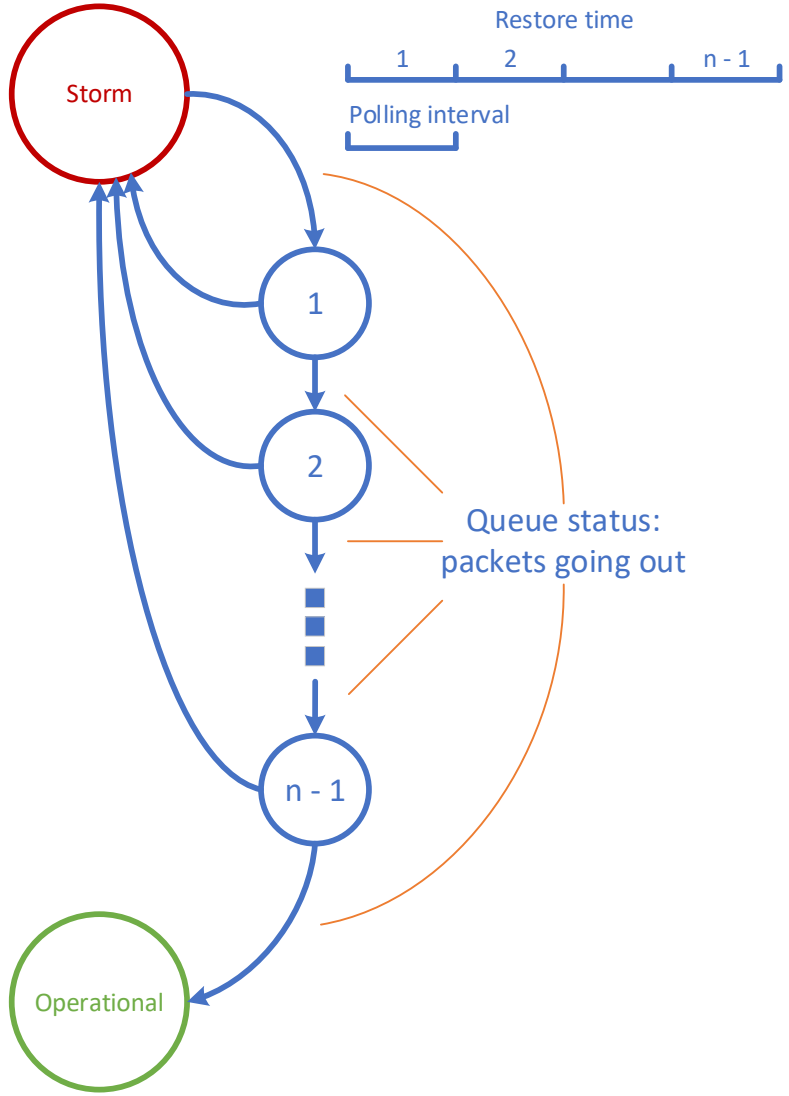
PFC storm detection logic

- Detection time
  - Time to deem the occurrence of a PFC storm
  - Generate a signal (to orchagent) to take drop actions
    - Tx packets are dropped
    - Rx packets are dropped



PFC storm restoration logic

- Restoration time
  - Time to deem the dismiss of a PFC storm
  - Generate a signal (to orchagent) to revoke drop actions

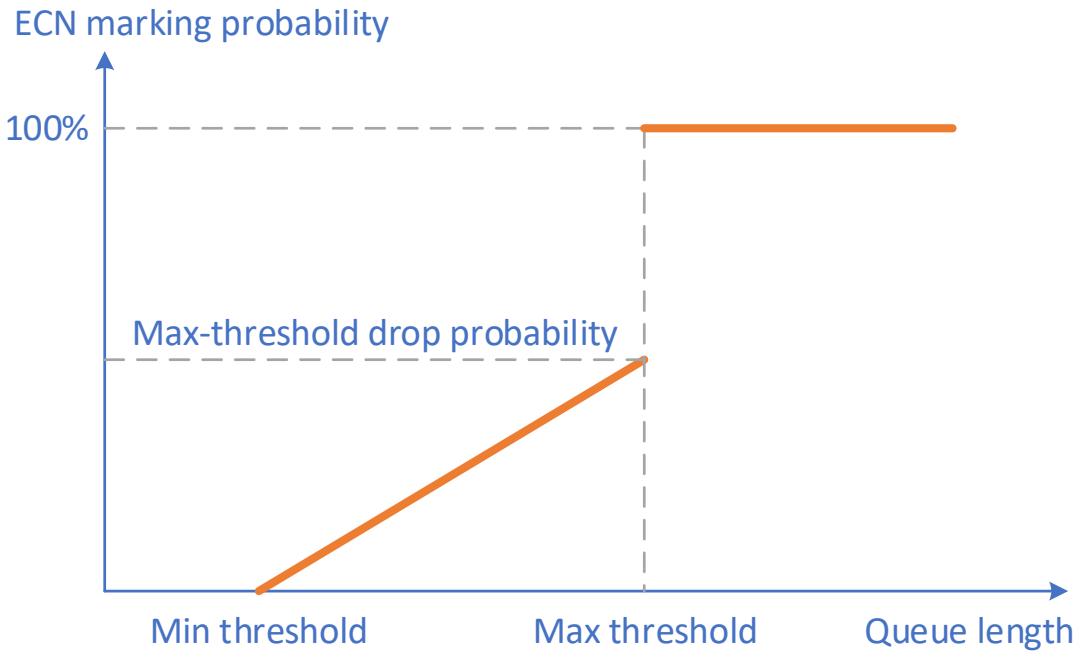


PFC watchdog configuration

```
admin@str-dx010-acs-1:~$ pfcwd show config
Changed polling interval to 200ms
       PORT    ACTION    DETECTION TIME    RESTORATION TIME
----------  --------  ----------------  ------------------
  Ethernet4      drop               200                 200
  Ethernet8      drop               200                 200
 Ethernet12      drop               200                 200
 Ethernet16      drop               200                 200
 Ethernet20      drop               200                 200
 Ethernet24      drop               200                 200
 Ethernet28      drop               200                 200
 Ethernet32      drop               200                 200
 Ethernet36      drop               200                 200
 Ethernet40      drop               200                 200
 Ethernet44      drop               200                 200
 Ethernet48      drop               200                 200
 Ethernet52      drop               200                 200
 Ethernet56      drop               200                 200
 Ethernet60      drop               200                 200
 Ethernet64      drop               200                 200
 Ethernet68      drop               200                 200
 Ethernet72      drop               200                 200
 Ethernet76      drop               200                 200
 Ethernet80      drop               200                 200
 Ethernet84      drop               200                 200
```

DCQCN support

- ECN marking




```
ecnconfig -l        ECN configuration

ecnconfig -q 3      ECN on/off status on queues
```

Roadmap

- RDMA config CLI
- Buffer pool watermark