RoCE (RDMA over Converged Ethernet) in Netberg SONiC for AI Workloads

Table of Contents

1. Supercharge Your Network for Lossless Traffic

The AI era has brought the demand for high-performance parallel computing to new heights. GPU clusters have become ubiquitous and reached a scale never seen before. Data-intensive workloads stress out GPUs and the communications between the GPU servers. The network performance became the backbone of the overall efficiency.

The network must provide seamless and reliable connectivity throughout the AI infrastructure, ensuring optimal performance for all AI training and inference tasks. RoCE (RDMA over Converged Ethernet) has become the key technology to removing communication bottlenecks and ensuring uninterrupted data flow between GPU servers. It allows efficient transfer of massive amounts of data between servers, keeping traffic lossless and uncongested.

SONiC has a strong QoS (Quality of Service) implementation that enables users to prioritize critical data traffic. This sends high-priority packets ahead of other traffic, and your important data is not lost.

You can turn your network into a lossless, low-latency, priority-driven fabric that easily handles your AI applications.

1.1. Making the Network Lossless And Prioritized

When packets enter the network fabric bearing DSCP/DOT1P marking, they can be mapped to queues on physical interfaces. Priority Flow Control (PFC) makes Ethernet lossless by generating hop-by-hop backpressure from the receiver toward the sender in the event of congestion. Rather than dropping packets, a pause frame is sent back to its sender, delaying transmission at the previous hop. Thus, PFC prevents packet drops for traffic of particular priority, achieving lossless flow.

Explicit Congestion Notification (ECN) is another technology to manage congestion. When the switch detects the congestion, the ECN field is set in the IP header of a packet before forwarding it. If the receiver device receives the packets with ECN marked, it sends the congestion notification packets (CNP) to the sender device to reduce the transmit packet rate. The CNP packets continue to be sent to the sender until the congestion is released.

These technologies ensure lossless priority traffic. Using queue scheduling, we can provide preferential treatment of traffic classes mapped to specific egress queues. This enhances performance even under congestion. SONiC supports three scheduling algorithms - Deficit Weighted Round Robin (DWRR), Weighted Round Robin (WRR), and Strict Priority Scheduling (SP). The system ensures that higher-priority traffic is transmitted preferentially, either in a weighted manner (for WRR/DWRR) or with absolute priority (for SP).

In conclusion, through the utilization of Priority Flow Control (PFC), Explicit Congestion Notification (ECN), and various scheduling techniques, SONiC guarantees that high-priority traffic from Graphics Processing Unit (GPU) servers is lossless and prioritized during periods of congestion on the egress stage.

1.2. How to Configure QoS?

Many companies claim that QoS in SONiC is hard to configure, even scary and daunting, and offer a management tool to handle that.

Such a tool makes you manually write the whole configuration, so you have to learn all the syntax and make no mistakes in the process.

But what does Netberg SONiC offer in this regard?

Configuring QoS in Netberg SONiC is as simple as this:

```
admin@sonic:~$ sudo config qos reload
```

What do you get with this?

Buffer profiles

Packet buffer management is an integral part of any QoS management (that almost no one talks about). Every switch has a shared packet buffer that can be split into pools for certain types of traffic. Such an approach guarantees that your high-priority lossless traffic will have a buffer area that other types of traffic cannot occupy.

After the profiles are created, another essential task is designating a buffer pool for traffic coming from a certain priority group of a port (queue). Priorities (queues) 3 and 4 are typically used for lossless traffic, while other priorities are used for lossy traffic.

The "config qos reload" command creates lossless and lossy pools and maps all seven priorities to these pools.

Traffic Classes and Queues

SONiC has default DSCP-to-TC Mapping (DSCP-to-TC mapping supersedes dot1p mapping, so the latter can be skipped https://netbergtw.com/top-support/netberg-sonic/qos-quality-of-service-classification/#_qos_classification).

DSCP	Traffic Class
0-2	1
3	3
4	4
5	2
6, 7	1
8	0
9-45	1
46	5
47	1
48	6

DSCP	Traffic Class
49-63	1

SONiC has default TC-to-Queue mapping that assigns the egress Queue ID based on the Traffic Class.

TC	0	1	2	3	4	5	6	7
Queue	0	1	2	3	4	5	6	7

SONiC has default TC-to-Priority-Group Mapping that maps the PFC priority to a priority group and enables the switch to set the PFC priority in xon/xoff frames in order to resume/pause corresponding egress queue at the link peer.

TC	0	1	2	3	4	5	6	7
PG	0	0	0	3	4	0	0	7

And SONiC binds these mapping values to all interfaces so nothing is amiss from the first try!

```
"Ethernet0": {
    "dscp_to_tc_map": "AZURE",
    "pfc_to_queue_map": "AZURE",
    "tc_to_pg_map": "AZURE",
    "tc_to_queue_map": "AZURE"
},
```

PFC and ECN

Even PFC (Priority Flow Control https://netbergtw.com/top-support/netberg-sonic/priority-flow-control-pfc/) and ECN (Explicit Congestion Notification https://netbergtw.com/top-support/netberg-sonic/wred-and-ecn/) are handled!

PFC is enabled on every intefaces queues 3 and 4.

```
"Ethernet0": {
    "dscp_to_tc_map": "AZURE",
    "pfc_enable": "3,4",
    "pfc_to_queue_map": "AZURE",
    "pfcwd_sw_enable": "3,4",
    "tc_to_pg_map": "AZURE",
    "tc_to_queue_map": "AZURE"
},
```

PFC is enabled on every interface queue 3 and 4.

An optional PFC Watchdog can be configured according to this piece https://netbergtw.com/top-support/netberg-sonic/pfc-priority-flow-control-watchdog/ PFC watchdog is designed to detect and mitigate PFC storm (abnormal back-pressure caused by receiving excessive PFC pause frames) received for each port.

A default WRED profile is created, and ECN uses its thresholds.

```
"WRED_PROFILE": {
    "AZURE_LOSSLESS": {
        "ecn": "ecn_all",
        "green_drop_probability": "5",
        "green_max_threshold": "2097152",
        "green_min_threshold": "1048576",
        "red_drop_probability": "5",
        "red_max_threshold": "2097152",
        "red_min_threshold": "1048576",
        "wred_green_enable": "true",
        "wred_red_enable": "true",
        "wred_yellow_enable": "true",
        "yellow_drop_probability": "5",
        "yellow_max_threshold": "2097152",
        "yellow_min_threshold": "1048576"
    }
}
```

Every PFC-enabled queue gets a WRED profile attached:

```
"Ethernet0|3": {
    "wred_profile": "AZURE_LOSSLESS"
},
"Ethernet0|4": {
    "wred_profile": "AZURE_LOSSLESS"
},
```

This default configuration proactively manages congestion to achieve lossless traffic flow on designated queues.

Optimizing Performance with Scheduling

Scheduling (https://netbergtw.com/top-support/netberg-sonic/qos-quality-of-service-classification/#_qos_scheduling) is critical to improving performance even under congestion. SONiC will generate two default schedulers:

```
"SCHEDULER": {
    "scheduler.0": {
        "type": "DWRR",
        "weight": "14"
    },
    "scheduler.1": {
        "type": "DWRR",
        "weight": "15"
    }
},
```

And attaches them to PFC-enabled queues.

```
"Ethernet0|0": {
    "scheduler": "scheduler.0"
},
"Ethernet0|1": {
```

```
    "scheduler": "scheduler.0"
},
"Ethernet0|2": {
    "scheduler": "scheduler.0"
},
"Ethernet0|3": {
    "scheduler": "scheduler.1",
    "wred_profile": "AZURE_LOSSLESS"
},
"Ethernet0|4": {
    "scheduler": "scheduler.1",
    "wred_profile": "AZURE_LOSSLESS"
},
```

Deficit Weighted Round Robin (DWRR) and Weighted Round Robin (WRR) scheduling operate with weights to determine a queue's share in the available bandwidth. The default configuration sets both lossless queues to equal treatment.

Summary

Netberg SONiC offers a default worry-free QoS configuration that requires only one command.

- Buffer profiles to separate lossy and lossless traffic.
- Each type of traffic gets routed to designated queues and managed appropriately.
- Queues 3 and 4 are Priority Flow Control (PFC) and Explicit Congestion Notification (ECN) pre-configured for lossless traffic and congestion management.
- Default schedulers give priority to lossless traffic.

1.3. Next Steps

The default configuration will serve you well at the beginning. As your network evolves, the QoS configuration might need some tuning to match new traffic patterns.

It's quite a simple task to edit the default configuration; just follow our guides! https://netbergtw.com/top-support/netberg-sonic/