

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 28, 2022

Z. Du
P. Liu
China Mobile
October 25, 2021

Micro-burst Decreasing in Layer3 Network for Low-Latency Traffic
draft-du-detnet-layer3-low-latency-04

Abstract

It is complex to support deterministic forwarding in a large scale network because there is too much dynamic traffic in the network and the data model becomes hard to predict after aggregation in the intermediate nodes. This document introduces the problem of micro-bursts in layer3 network, and analyses the method to decrease the micro-bursts in layer3 network for low-latency traffic.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 28, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

(<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Gaps for Large-scale Layer 3 Deterministic Network	3
3. Micro-burst Problem in IP Forwarding	3
4. Analysis of the Method to Decrease Micro-bursts	5
5. An Example of Method to Decrease Micro-bursts	5
5.1. Working Flow of the Method	6
5.2. Process of Edge Node	6
5.3. Process of Forwarding Node	6
5.4. Analysis of the Proposed Method	7
6. IANA Considerations	8
7. Security Considerations	8
8. Acknowledgements	8
9. References	8
9.1. Normative References	8
9.2. Informative References	8
Authors' Addresses	9

1. Introduction

The DetNet architecture [RFC8655] is supposed to work in campus-wide networks and private WANs, including the large-scale ISP network scenario, such as the 5G bearing network, as mentioned in [RFC8578]. It is essential for the large-scale ISP network to be able to provide the low-latency service. The low-latency requirement exists in both L2 and L3 networks, and in both small and large networks.

However, as talked in [I-D.qiang-detnet-large-scale-detnet], deploying deterministic services in a large-scale network brings a lot of new challenges. A novel method called LDN (Large-scale Deterministic Network) is introduced in [I-D.qiang-detnet-large-scale-detnet] and [I-D.dang-queuing-with-multiple-cyclic-buffers], which explore the deterministic forwarding over a large-scale network.

This document also explores the deterministic service in the large-scale layer 3 network, and analyses the method based on micro-burst decreasing, which can benefit the forwarding of low-latency traffic in the large-scale network.

2. Gaps for Large-scale Layer 3 Deterministic Network

In this document, the large-scale network means that there are many dynamic flows in the network, but it is hard to do per-flow shaping on the intermediate nodes because they have high pressure on forwarding on the data plane.

According to [RFC8655], DetNet operates at the IP layer and delivers service over lower-layer technologies such as MPLS and IEEE 802.1 Time-Sensitive Networking [TSN]. However, the TSN mechanisms are designed for L2 network originally, and cannot be directly used in the large-scale layer 3 network because of various reasons. Some of them are described as below.

Some TSN mechanisms need synchronization of the network equipments, which is easier in a small network, but hard in a large network. It brings in some complex maintenance jobs across a long distance that are not needed before.

Some TSN mechanisms need a per-flow state in the forwarding plane, which is un-scalable. Aggregation methods need to be considered.

Some TSN mechanisms need a constant and forecastable traffic characteristics, which is more complicated in a large network which includes much more flows joining in or leaving randomly and the traffic characteristics are more dynamic.

The main aspects of the problems are the simplicity and the scalability. The former can ensure that the mechanism is easy to deploy, and the second can ensure that the mechanism is able to bear a large number of deterministic services.

3. Micro-burst Problem in IP Forwarding

The current IP forwarding mechanism is considered to be a good example fulfilling the requirements of simplicity and scalability. However, the traditional IP network is based on statistical multiplexing, and can only provide Best Effort service, short of SLA guaranteed mechanisms.

When we rethink the problem in the current IP forwarding mechanism, we can find that in the current IP network, a long delay in queuing, or some packet losses due to burst are acceptable; however, it may be unacceptable in the deterministic forwarding. Therefore, they have different design principles in the low layer.

The current forwarding mechanism in an IP router, which is based on statistical multiplexing, can not provide the deterministic service

because of various reasons. Even be given a high priority, a critical packet can experience a long congestion delay or be lost in a relatively light-loaded network, which is caused by micro-bursts in the network.

Micro-burst is a special case of network congestion, which typically lasts a short period, at the granularity of millisecond. In a micro-burst, a lot of data are received on the interface suddenly, and the temporary bandwidth requirement would be tens of or hundreds of the average bandwidth requirement, or even exceed the interface bandwidth.

In most cases, the buffer on the equipment can handle the micro-bursts. However, in some corner cases, micro-bursts bring in a long delay (for example, at the granularity of millisecond) or even packet loss.

The following paragraphs introduce the causes of the micro-burst.

Firstly, IP traffic has an instinct of burstiness no matter in the macro or micro aspect, i.e., it does not have a constant traffic model even after aggregations.

Secondly, IP network has a flexible topology, where the incoming traffic may exceed the bandwidth of the outgoing interface. For example, an interface with a large bandwidth may need to send traffic to an interface with a smaller bandwidth, or multiple flows from several incoming interfaces may need to occupy the same outgoing interface.

Thirdly, the IP node has been designed to send traffic as quickly as possible, and it is not aware whether the downstream node's buffer can handle the traffic. For example, Figure 1 below shows the problem of the current IP scheduling mechanism. Before the scheduling in an IP network, the packets are well paced, but after the scheduling, the packets will be gathered even the total traffic rate is unchanged. When an IP outgoing interface receives multiple critical flows from several incoming interfaces, the situation becomes worse. However, an IP router will try to send them as soon as possible, so occasionally, in some later hops, micro-bursts will emerge.

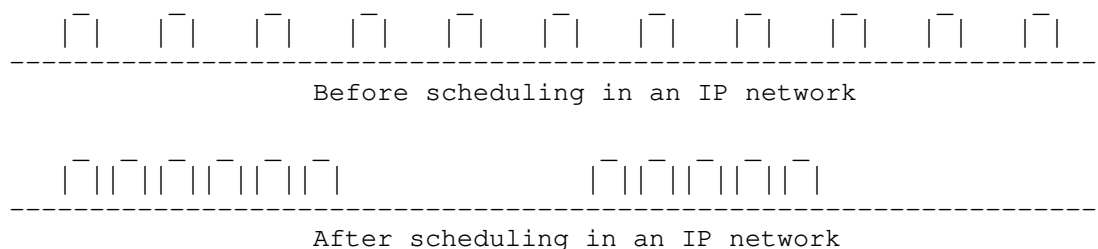


Figure 1: Change of the traffic characteristics in an IP network

4. Analysis of the Method to Decrease Micro-bursts

This document analyses the method to support the low latency traffic bearing in an IP network, such as the 5G bearing network, by avoiding micro-bursts in the network as much as possible. The principle in this method is to forward critical and BE traffic separately, and do not distinguish different critical flows on the forwarding plane on the intermediate nodes.

As talked before, the target method should be scalable and easy to deploy. As the intermediate nodes have high pressure on forwarding packets, the target method should not bring in too much complex process on the data plane. Several requirements are listed as follows.

The first is that the DetNet traffic should support aggregation. The intermediate nodes should not do per-flow process on the data plane.

The second is that separation process of the control plane and data plane on the intermediate nodes. The status of the aggregated DetNet traffic on the control plane may change frequently in the large-scale network. We should not assume that the control plane on an intermediate node can interact with the data plane frequently, for example, to change a shaper parameter frequently. On the data plane, some self-decision process should be supported.

5. An Example of Method to Decrease Micro-bursts

In this section, we describes an example of method fulfilling the requirements mentioned in the last section. It needs the cooperation of the edge nodes and the forwarding/intermediate nodes in an IP network.

5.1. Working Flow of the Method

Generally, the method contains two steps:

Step1: per flow schedule on the edge node. The purpose is to make sure that each critical traffic has a constant traffic model.

Step2: per interface schedule on the intermediate node. Traffic are aggregated to ensure the scalability, and the pacing also makes sure that they do not gather. The purpose is to make the critical traffic be forwarded as the shape when outgoing the edge, not as quickly as possible. We assume that the sending rate of the buffer for the critical traffic is the same as the receiving rate (how to achieve this is out of scope of this document). If all work well, the buffer will be maintained with a proper depth.

Other requirements include an RSVP-TE liked mechanism with a good scalability, which should be used to make sure the bandwidth is not exceeded on the interface.

5.2. Process of Edge Node

The edge node of the IP network can recognize each critical flows just as in the TSN network, and then give them individually a good shaping. In fact, in TSN mechanisms, no micro-burst will emerge for critical traffic, and each TSN mechanism is proved to be effective under certain conditions.

This document suggests the edge node to shape the critical traffic by using the CBS method in [IEEE802.1Qav], or the shaping methods in [IEEE802.1Qcr]. Generally, the shaping methods can generate a paced traffic for each critical flow.

The parameters of the shaper, such as the sending rate, can be configured for each flow by some means.

5.3. Process of Forwarding Node

For the forwarding node, it is uneasy to recognize each critical flow because of the high pressure of forwarding a large amount of packets. It is suggested that no per-flow state is maintained on the forwarding node. It is to say that, on the forwarding node, the critical flows should be aggregated and handled together.

This document suggests that the forwarding node can deploy a specific queue on each outgoing interface. The queue will buffer all critical traffic that need to go out through that interface, and will pace them by using methods mentioned in the last section.

A shaping method in TSN is used here instead of the original forwarding method in an IP router, which can make the critical traffic be forwarded orderly instead of as soon as possible. Therefore, micro-bursts can be decreased in the network.

If all the forwarding nodes can do their jobs properly, i.e., they can well pace the critical traffic, no or rare micro-bursts for the critical traffic would take place. In this way, the critical traffic will have a relatively low latency in the IP network with less uncertainty of micro-bursts.

As no per-flow state is maintained on the forwarding node, the sending rate of the shaper is hard to decide. As said in the last session, the sending rate is suggested to be adjusted referring to the incoming rate of the queue. The purpose is to maintain a proper buffer depth for the queue.

Although it is claimed that the proposed method is simpler than the TSN mechanisms, forwarding/intermediate nodes also need to be updated. The detailed realization of the method on the intermediate nodes is out of scope of this document.

5.4. Analysis of the Proposed Method

The method proposed does not need synchronization, just as the asynchronous mechanisms studied in [IEEE802.1Qcr]. Furthermore, the method has a larger aggregation granularity, which can fulfill the requirements of simplicity and scalability as much as possible. However, in theory, it has a larger uncertainty on the forwarding than the zero congestion loss target in the TSN mechanisms.

We compare three mechanisms in the following paragraphs. The first is the priority based light-load mechanism, i.e., the traditional method. The second is the TSN mechanism, such as CQF. The third is the proposed mechanism.

In the first mechanism, we only give a high priority to the critical traffic, and thus the scalability of the deterministic system is good. However, the uncertainty on the forwarding plane perhaps can not fulfill the requirements in the industry network where SLA requirements are very essential. Perhaps, it is only able to work well when a small amount of critical traffic exist in the network.

If we use the scheduling method in the TSN, such as CQF. Its uncertainty is very low, but its scalability is not very good as said in Section 2. It should be noted that in a large deterministic system, the ISP normally will not guarantee the user 100 percent reliability, instead of which it perhaps is a value very close to.

The proposed method has a better scalability than the TSN mechanisms, and a better reliability than the priority based method. If we assume that different services need different deterministic levels, this method may be helpful for the service that does not need a very high deterministic level. For example, the method can be used in the consumption Internet, in which the deterministic service needs a relatively lower deterministic level than the industry Internet.

6. IANA Considerations

This document has no IANA actions.

7. Security Considerations

Detailed security considerations can refer to [I-D.ietf-detnet-bounded-latency] and [I-D.ietf-detnet-security].

8. Acknowledgements

Thanks for the valuable comments from Janos Farkas, Lou Berger, and David Black.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

[RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.

9.2. Informative References

[I-D.dang-queuing-with-multiple-cyclic-buffers] Liu, B. and J. Dang, "A Queuing Mechanism with Multiple Cyclic Buffers", draft-dang-queuing-with-multiple-cyclic-buffers-00 (work in progress), February 2021.

[I-D.ietf-detnet-bounded-latency] Finn, N., Boudec, J. L., Mohammadpour, E., Zhang, J., Varga, B., and J. Farkas, "DetNet Bounded Latency", draft-ietf-detnet-bounded-latency-07 (work in progress), September 2021.

- [I-D.ietf-detnet-security]
Grossman, E., Mizrahi, T., and A. J. Hacker,
"Deterministic Networking (DetNet) Security
Considerations", draft-ietf-detnet-security-16 (work in
progress), March 2021.
- [I-D.qiang-detnet-large-scale-detnet]
Qiang, L., Geng, X., Liu, B., Eckert, T., Geng, L., and G.
Li, "Large-Scale Deterministic IP Network", draft-qiang-
detnet-large-scale-detnet-05 (work in progress), September
2019.
- [IEEE802.1Qav]
IEEE 802.1, "IEEE 802.1Qav-2009 - IEEE Standard for Local
and metropolitan area networks-- Virtual Bridged Local
Area Networks Amendment 12: Forwarding and Queuing
Enhancements for Time-Sensitive Streams", 2009,
<https://standards.ieee.org/standard/802_1Qav-2009.html>.
- [IEEE802.1Qcr]
IEEE 802.1, "IEEE 802.1Qcr-2020 - IEEE Standard for Local
and Metropolitan Area Networks--Bridges and Bridged
Networks - Amendment 34: Asynchronous Traffic Shaping",
2020,
<https://standards.ieee.org/standard/802_1Qcr-2020.html>.
- [RFC8578] Grossman, E., Ed., "Deterministic Networking Use Cases",
RFC 8578, DOI 10.17487/RFC8578, May 2019,
<<https://www.rfc-editor.org/info/rfc8578>>.
- [TSN] IEEE 802.1, "Time-Sensitive Networking (TSN) Task Group",
2012, <<https://1.ieee802.org/tsn/>>.

Authors' Addresses

Zongpeng Du
China Mobile
No.32 XuanWuMen West Street
Beijing 100053
China

Email: duzongpeng@foxmail.com

Peng Liu
China Mobile
No.32 XuanWuMen West Street
Beijing 100053
China

Email: liupengyjy@chinamobile.com

DETNET
Internet-Draft
Intended status: Standards Track
Expires: 28 April 2022

T. Eckert
Futurewei Technologies USA
S. Bryant
University of Surrey ICS
A.G. Malis
Malis Consulting
25 October 2021

Deterministic Networking (DetNet) Data Plane - MPLS TC Tagging for
Cyclic Queuing and Forwarding (MPLS-TC TCQF)
draft-eckert-detnet-mpls-tc-tcqf-01

Abstract

This memo defines the use of the MPLS TC field of MPLS Label Stack Entries (LSE) to support cycle tagging of packets for Multiple Buffer Cyclic Queuing and Forwarding (TCQF). TCQF is a mechanism to support bounded latency forwarding in DetNet network.

Target benefits of TCQF include low end-to-end jitter, ease of high-speed hardware implementation, optional ability to support large number of flow in large networks via DiffServ style aggregation by applying TCQF to the DetNet aggregate instead of each DetNet flow individually, and support of wide-area DetNet networks with arbitrary link latencies and latency variations.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 28 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction (informative)	2
2. Using TCWF in the DetNet Architecture and MPLS forwarding plane (informative)	3
3. MPLS T-CWF forwarding (normative)	6
3.1. Configuration Data model and tag processing for MPLS TC TCWF	6
3.2. Packet processing	6
3.3. TCWF with label stack operations	7
3.4. Ingress operations	8
4. TCWF Pseudocode (normative)	8
5. Operational considerations (informative)	9
5.1. Controller plane computation of cycle mappings	10
6. Security Considerations	11
7. IANA Considerations	11
8. Changelog	11
9. References	12
9.1. Normative References	12
9.2. Informative References	12
Authors' Addresses	13

1. Introduction (informative)

Cyclic Queuing and Forwarding [CWF], is an IEEE standardized queuing mechanism in support of deterministic bounded latency. See also [I-D.ietf-detnet-bounded-latency], Section 6.6.

CWF benefits for Deterministic QoS include the tightly bounded jitter it provides as well as the per-flow stateless operation, minimizing the complexity of high-speed hardware implementations and allowing to support on transit hops arbitrary number of DetNet flow in the forwarding plane because of the absence of per-hop, per-flow QoS processing. In the terms of the IETF QoS architecture, CWF can be called DiffServ QoS technology, operating only on a traffic aggregate.

CQFs is limited to only limited-scale wide-area network deployments because it cannot take the propagation latency of links into account, nor potential variations thereof. It also requires very high precision clock synchronization, which is uncommon in wide-area network equipment beyond mobile network fronthaul. See [I-D.eckert-detnet-bounded-latency-problems] for more details.

This specification introduces and utilizes an enhanced form of CQF where packets are tagged with a cycle identifier, and a limited number of cycles, e.g.: 3...7 are used to overcome these distance and clock synchronization limitations. Because this memo defines how to use the TC field of MPLS LSE as the tag to carry the cycle identifier, it calls this scheme TC Tagged multiple buffer CQF (TC TCQF). See [I-D.qiang-DetNet-large-scale-DetNet] and [I-D.dang-queuing-with-multiple-cyclic-buffers] for more details of the theory of operations of TCQF. Note that TCQF is not necessarily limited to deterministic operations but could also be used in conjunction with congestion controlled traffic, but those considerations are outside the scope of this memo.

TCQF is likely especially beneficial when MPLS networks are designed to avoid per-hop, per-flow state even for traffic steering, which is the case for networks using SR-MPLS [RFC8402] for traffic steering of MPLS unicast traffic and/or BIER-TE [I-D.ietf-bier-te-arch] for tree engineering of MPLS multicast traffic. In these networks, it is specifically undesirable to require per-flow signaling to P-LSR solely for DetNet QoS because such per-flow state is unnecessary for traffic steering and would only be required for the bounded latency QoS mechanism and require likely even more complex hardware and manageability support than what was previously required for per-hop steering state (e.g. In RSVP-TE). Note that the DetNet architecture [RFC8655] does not include full support for this DiffServ model, which is why this memo describes how to use MPLS TC TCQF with the DetNet architecture per-hop, per-flow processing as well as without it.

2. Using TCQF in the DetNet Architecture and MPLS forwarding plane (informative)

This section gives an overview of how the operations of T-CQF relates to the DetNet architecture. We first revisit QoS with DetNet in the absence of T-CQF.

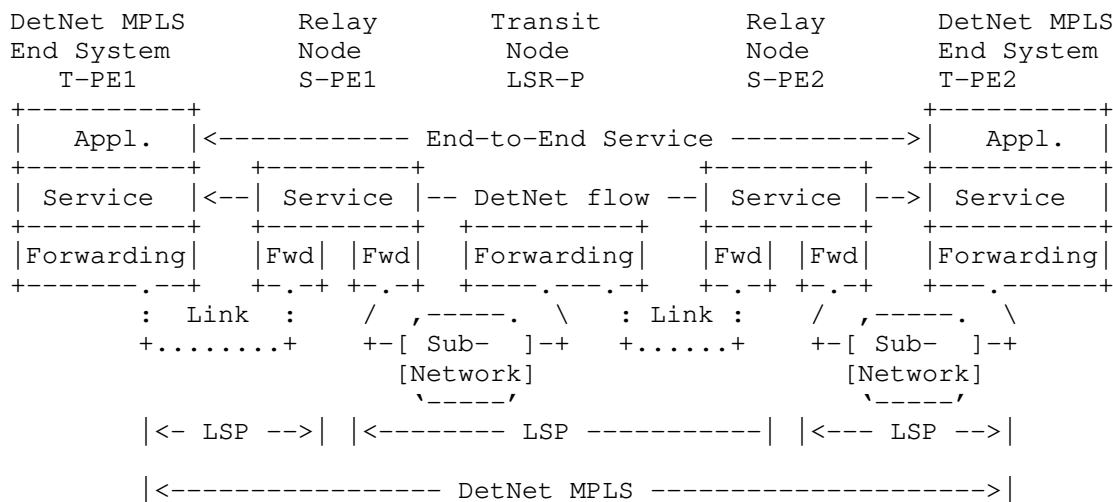


Figure 1: A DetNet MPLS Network

The above Figure 1, is copied from [RFC8964], Figure 2, and only enhanced by numbering the nodes to be able to better refer to them in the following text.

Assume a DetNet flow is sent from T-PE1 to T-PE2 across S-PE1, LSR, S-PE2. In general, bounded latency QoS processing is then required on the outgoing interface of T-PE1 towards S-PE1, and any further outgoing interface along the path. When T-PE1 and S-PE2 know that their next-hop is a service LSR, their DetNet flow label stack may simply have the DetNet flows Service Label (S-Label) as its Top of Stack (ToS) LSE, explicitly indicating one DetNet flow.

On S-PE1, the next-hop LSR is not DetNet aware, which is why S-PE1 would need to send a label stack where the S-Label is followed by a Forwarding Label (F-Label), and LSR-P would need to perform bounded latency based QoS on that F-Label.

For bounded latency QoS mechanisms relying on per-flow regulator state, such as in [TSN-ATS], this requires the use of a per-detnet flow F-Label across the network from S-PE1 to S-PE2, for example through RSVP-TE [RFC3209] enhanced as necessary with QoS parameters matching the underlying bounded latency mechanism (such as [TSN-ATS]).

With TC TCWF, a sequence of LSR and DetNet service node implements TC TCWF, ideally from T-PE1 (ingress) to T-PE2 (egress). The ingress node needs to perform per-DetNet-flow per-packet "shaping" to assign each packet of a flow to a particular TCWF cycle. This ingress-edge-function is currently out of scope of this document (TBD), but would be based on the same type of edge function as used in CWF.

All LSR/Service node after the ingress node only have to map a received TCWF tagged DetNet packet to the configured cycle on the output interface, not requiring any per-DetNet-flow QoS state. These LSR/Service nodes do therefore also not require per-flow interactions with the controller plane for the purpose of bounded latency.

Per-flow state therefore is therefore only required on nodes that are DetNet service nodes, or when explicit, per-DetNet flow steering state is desired, instead of ingress steering through e.g.: SR-MPLS.

Operating TCWF per-flow stateless across a service node, such as S-PE1, S-PE2 in the picture is only an option. It is of course equally feasible to Have one TCWF domain from T-PE1 to S-PE2, start a new TCWF domain there, running for example up to S-PE2 and start another one to T-PE2.

A service node must act as an egress/ingress edge of a TCWF domain if it needs to perform operations that do change the timing of packets other than the type of latency that can be considered in configuration of TCWF (see Section 5.1).

For example, if T-PE1 is ingress for a TCWF domain, and T-PE2 is the egress, S-PE1 could perform the DetNet Packet Replication Function (PRF) without having to be a TCWF edge node as long as it does not introduce latencies not included in the TCWF setup and the controller plane reserves resources for the multitude of flows created by the replication taking the allocation of resources in the TCWF cycles into account.

Likewise, S-PE2 could perform the Packet Elimination Function without being a TCWF edge node as this most likely does not introduce any non-TCWF acceptable latency - and the controller plane accordingly reserves only for one flow the resources on the S-PE2->T-PE2 leg.

If on the other hand, S-PE2 was to perform the Packet Reordering Function (PRF), this could create large peaks of packets when out-of-order packets are released together. A PRF would either have to take care of shaping out those bursts for the traffic of a flow to again conform to the admitted CIR/PIR, or else the service node would have to be a TCWF egress/ingress, performing that shaping itself as an ingress function.

3. MPLS T-CWF forwarding (normative)

3.1. Configuration Data model and tag processing for MPLS TC TCWF

```

tcwf
+-- uint16 cycles
+-- uint16 cycle_time
+-- uint32 cycle_clock_offset
+-- if_config[oif] # Outgoing InterFace
    +-- uint32 cycle_clock_offset
    +-- cycle_map[iif] # Incoming InterFace
        +--uint8 oif_cycle[iif_cycle]

tcwf_tc[oif]
+--uint8 tc[oif_cycle]
```

Figure 2: TCWF Configuration Data Model

3.2. Packet processing

This section explains the MPLS T-CWF packet processing and through it, introduces the semantic of the objects in Figure 2

tcwf contains the router/LSR wide configuration of TCWF parameters, independent of the specific tagging mechanism on any interface. Any interface can have a different tagging method.

The model represents a single TCWF domain, which is a set of interfaces acting both as ingress (iif) and egress (oif) interfaces, capable to forward TCWF packets amongst each other. A router/LSR may have multiple TCWF domains each with a set of interfaces disjoint from those of any other TCWF domain.

tcwf.cycles is the number of cycles used across all interfaces in the TCWF domain. router/LSR MUST support 3 and 4 cycles. To support interfaces with MPLS TC tagging, 7 or less cycles MUST be used across all interfaces in the CWF domain.

The unit of tcwf.cycle_time is micro-seconds. router/LSR MUST support configuration of cycle-times of 20,50,100,200,500,1000,2000 usec.

Cycles start at an offset of tcwf.cycle_clock_offset in units of nsec as follows. Let clock1 be a timestamp of the local reference clock for TCWF, at which cycle 1 starts, then:

```

tcwf.cycle_clock_offset = (clock1 mod (tcwf.cycle_time * tcwf.cycles)
)
```


The local reference clock of the LSR/router is expected to be synchronized with the neighboring LSR/router in TCQF domain. `tcqw.cycle_clock_offset` can be configurable by the operator, or it can be read-only. In either case will the operator be able to configure working TCQF forwarding through appropriately calculated cycle mapping.

`tcqw.if_config[oif]` is optional per-interface configuration of TCQF parameters. `tcqw.if_config[oif].cycle_clock_offset` may be different from `tcqw.cycle_clock_offset`, for example, when interfaces are on line cards with independently synchronized clocks, or when non-uniform ingress-to-egress propagation latency over a complex router/LSR fabric makes it beneficial to allow per-egress interface or line card configuration of `cycle_clock_offset`. It may be configurable or read-only.

The value of -1 for `tcqw.if_config[oif].cycle_clock_offset` is used to indicate that the domain wide `tcqw.cycle_clock_offset` is to be used for `oif`. This is the only permitted negative number for this parameter.

When a packet is received from `iif` with a cycle value of `iif_cycle` and the packet is routed towards `oif`, then the cycle value (and buffer) to use on `oif` is `tcqw.if_config[oif].cycle_map[iif].oif_cycle[iif_cycle]`. This is called the cycle mapping and is must be configurable. This cycle mapping always happens when the packet is received with a cycle tag on an interface in a TCQF domain and forwarded to another interface in the same TCQF domain.

`tcqw_tc[oif].tc[oif_cycle]` defines how to map from the internal cycle number `oif_cycle` to an MPLS TC value on interface `oif`. When `tcqw_tc[oif]` is configured, `oif` will use MPLS TC tagging for TCQF. This mapping not only used to map from internal cycle number to MPLS TC tag when sending packets, but also to map from MPLS TC tag to the internal cycle number when receiving packets.

3.3. TCQF with label stack operations

In the terminology of [RFC3270], TCQF QoS as defined here, is TC-Inferred-PSC LSP (E-LSP) behavior: Packets are determined to belong to the TCQF PSC solely based on the TC of the received packet.

The internal cycle number SHOULD be assigned from the Top of Stack (ToS) MPLS label TC bits before any other label stack operations happens. On the egress side, the TC value of the ToS MPLS label SHOULD be assigned from the internal cycle number after any label stack processing.

With this order of processing, TCWF can support forwarding of packets with any label stack operations such as label swap in the case of LDP or RSVP-TE created LSP, or no label changes from SID hop-by-hop forwarding and/or SID/label pop as in the case of SR-MPLS traffic steering.

3.4. Ingress operations

The ingress LSR of a TCWF domain has to mark packets with an internal cycle number and ensure that any such marked traffic complies with the traffic envelope admitted by the controller plane.

The algorithms to map packets of traffic flows into cycles are outside the scope of this specification, because there can be multiple ones of varying complexity. In a most simple admission model, a particular flow is allocated a maximum number of bytes in every cycle. This can easily be mapped into an appropriate policing gate.

For the purpose of this specification, such ingress operations is simply represented as an (internal/virtual) interface from which the packet is received, complete with a correctly assigned internal cycle number.

4. TCWF Pseudocode (normative)

The following pseudocode restates the forwarding behavior of Section 3 in an algorithmic fashion as pseudocode. It uses the objects of the TCWF configuration data model defined in Section 3.1.

```
void receive(pak) {
    // Receive side TCWF - remember cycle in
    // packet internal header
    iif = pak.context.iif
    if (tcwf.if_config[iif]) { // TCWF enabled on iif
        if (tcwf_tc[iif]) { // MPLS TCWF enabled on iif
            tc = pak.mpls_header.lse[ tos ].tc
            pak.context.tcwf_cycle = map_tc2cycle( tc, tcwf_tc[iif] )
        } else // other future encap/tagging options for TCWF
        {
        }
    }
    forward(pak);
}

void inject_tcwf_pak(pak, cycle) {
    pak.context.iif = INTERNAL
    pak.context.tcwf_cycle = cycle
    forward(pak);
}
```

```

void forward(pak) {
    // Forwarding including any LSE operations
    oif = pak.context.oif = forward_process(pak)

    // ... optional DetNet PREOF functions here
    // ... if router is DetNet service node

    if(pak.context.tcqv_cycle && // non TCQF packets cycle is 0
        tcqv.if_config[oif]) { // TCQF enabled
        // Map tcqv_cycle iif to oif
        cycle = pak.context.tcqv_cycle
            = map_cycle(cycle,
                tcqv.if_config[oif].cycle_map[[iif])

        if(tcqv.mpls_tc_tag[iif]) { // TC-TCQF
            pak.mpls_header.lse[tos].tc =
                map_cycle2tc(cycle, tcqv_tc[oif])
        } else // other future encap/tagging options for TCQF

            tcqv_enqueue(pak, oif.cycleq[cycle])
    }
}

// Started when TCQF is enabled on an interface
// dequeues packets from oif.cycleq
void send_tcqv(oif) {
    cycle = 1
    cc = tcqv.cycle_time *
        tcqv.cycle_time
    o = tcqv.cycle_clock_offset
    nextcyclestart = floor(tnow / cc) * cc + cc + o

    while(1) {
        while(tnow < nextcyclestart) { }
        while(pak = dequeue(oif.cycleq(cycle))) {
            send(pak)
        }
        cycle = (cycle + 1) mod tcqv.cycles + 1
        nextcyclestart += tcqv.cycle_time
    }
}

```

Figure 3: TCQF Pseudocode

5. Operational considerations (informative)

5.1. Controller plane computation of cycle mappings

The cycle mapping is computed by the controller plane by taking at minimum the link, interface serialization and node internal forwarding latencies as well as the cycle_clock_offsets into account.

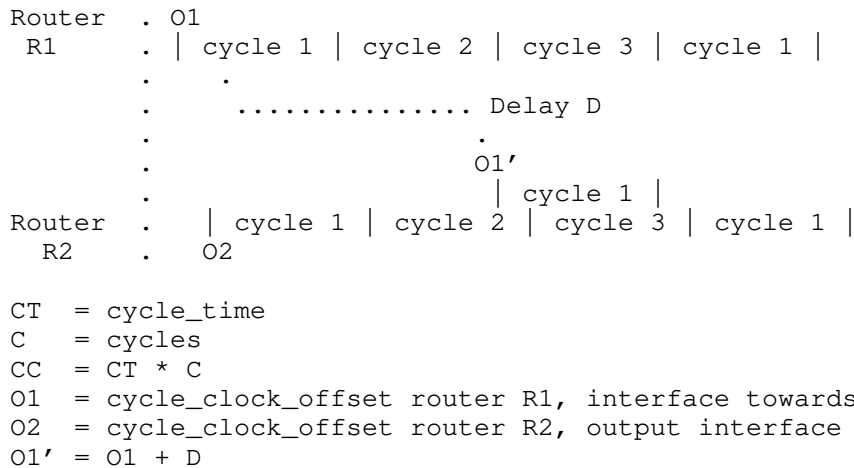


Figure 4: Calculation reference

Consider in {#Calc1} that Router R1 sends packets via $C = 3$ cycles with a cycle_clock offset of $O1$ towards Router R2. These packets arrive at R2 with a cycle_clock offset of $O1'$ which includes through D all latencies incurred between releasing a packet on R1 from the cycle buffer until it can be put into a cycle buffer on R2: serialization delay on R1, link delay, non_CQF delays in R1 and R2, especially forwarding in R2, potentially across an internal fabric to the output interface with the sending cycle buffers.

$$A = (\text{ceil}((O1' - O2) / CT) + C + 1) \bmod CC$$

$$\text{map}(i) = (i - 1 + A) \bmod C + 1$$

Figure 5: Calculating cycle mapping

{#Calc2} shows a formula to calculate the cycle mapping between R1 and R2, using the first available cycle on R2. In the example of {#Calc1} with $CT = 1$, $(O1' - O2) = \sim 1.8$, A will be 0, resulting in $\text{map}(1)$ to be 1, $\text{map}(2)$ to be 2 and $\text{map}(3)$ to be 3.

The offset "C" for the calculation of A is included so that a negative $(O1 - O2)$ will still lead to a positive A .

In general, D will be variable [$D_{min}...D_{max}$], for example because of differences in serialization latency between min and max size packets, variable link latency because of temperature based length variations, link-layer variability (radio links) or in-router processing variability. In addition, D also needs to account for the drift between the synchronized clocks for $R1$ and $R2$. This is called the Maximum Time Interval Error (MTIE).

Let $A(d)$ be A where $O1'$ is calculated with $D = d$. To account for the variability of latency and clock synchronization, $map(i)$ has to be calculated with $A(D_{max})$, and the controller plane needs to ensure that that $A(D_{min})...A(D_{max})$ does cover at most $(C - 1)$ cycles.

If it does cover C cycles, then C and/or CT are chosen too small, and the controller plane needs to use larger numbers for either.

This $(C - 1)$ limitation is based on the understanding that there is only one buffer for each cycle, so a cycle cannot receive packets when it is sending packets. While this could be changed by using double buffers, this would create additional implementation complexity and not solve the limitation for all cases, because the number of cycles to cover [$D_{min}...D_{max}$] could also be $(C + 1)$ or larger, in which case a tag of $1...C$ would not suffice.

6. Security Considerations

TBD.

7. IANA Considerations

This document has no IANA considerations.

8. Changelog

00

Initial version

01

Added new co-author.

Changed Data Model to "Configuration Data Model",

and changed syntax from YANG tree to a non-YANG tree, removed empty section targeted for YANG model. Reason: the configuration parameters that we need to specify the forwarding behavior is only a subset of what likely would be a good YANG model, and any work to

define such a YANG model not necessary to specify the algorithm would be scope creep for this specification. Better done in a separate YANG document. Example additional YANG aspects for such a document are how to map parameters to configuration/operational space, what additional operational/monitoring parameter to support and how to map the YANG objects required into various pre-existing YANG trees.

Improved text in forwarding section, simplified sentences, used simplified configuration data model.

9. References

9.1. Normative References

- [RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, DOI 10.17487/RFC3270, May 2002, <<https://www.rfc-editor.org/info/rfc3270>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.
- [RFC8964] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., Bryant, S., and J. Korhonen, "Deterministic Networking (DetNet) Data Plane: MPLS", RFC 8964, DOI 10.17487/RFC8964, January 2021, <<https://www.rfc-editor.org/info/rfc8964>>.

9.2. Informative References

- [CQF] IEEE Time-Sensitive Networking (TSN) Task Group., "IEEE Std 802.1Qch-2017: IEEE Standard for Local and Metropolitan Area Networks - Bridges and Bridged Networks - Amendment 29: Cyclic Queuing and Forwarding", 2017.
- [I-D.dang-queuing-with-multiple-cyclic-buffers] Liu, B. and J. Dang, "A Queuing Mechanism with Multiple Cyclic Buffers", Work in Progress, Internet-Draft, draft-dang-queuing-with-multiple-cyclic-buffers-00, 22 February 2021, <<https://www.ietf.org/archive/id/draft-dang-queuing-with-multiple-cyclic-buffers-00.txt>>.
- [I-D.eckert-detnet-bounded-latency-problems] Eckert, T. and S. Bryant, "Problems with existing DetNet bounded latency queuing mechanisms", Work in Progress, Internet-Draft, draft-eckert-detnet-bounded-latency-

problems-00, 12 July 2021,
<<https://www.ietf.org/archive/id/draft-eckert-detnet-bounded-latency-problems-00.txt>>.

[I-D.ietf-bier-te-arch]

Eckert, T., Cauchie, G., and M. Menth, "Tree Engineering for Bit Index Explicit Replication (BIER-TE)", Work in Progress, Internet-Draft, draft-ietf-bier-te-arch-10, 9 July 2021, <<https://www.ietf.org/archive/id/draft-ietf-bier-te-arch-10.txt>>.

[I-D.ietf-detnet-bounded-latency]

Finn, N., Boudec, J. L., Mohammadpour, E., Zhang, J., Varga, B., and J. Farkas, "DetNet Bounded Latency", Work in Progress, Internet-Draft, draft-ietf-detnet-bounded-latency-07, 1 September 2021, <<https://www.ietf.org/archive/id/draft-ietf-detnet-bounded-latency-07.txt>>.

[I-D.qiang-DetNet-large-scale-DetNet]

Qiang, L., Geng, X., Liu, B., Eckert, T., Geng, L., and G. Li, "Large-Scale Deterministic IP Network", Work in Progress, Internet-Draft, draft-qiang-DetNet-large-scale-DetNet-05, 2 September 2019, <<https://www.ietf.org/archive/id/draft-qiang-DetNet-large-scale-DetNet-05.txt>>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.

[TSN-ATS] Specht, J., "P802.1Qcr - Bridges and Bridged Networks Amendment: Asynchronous Traffic Shaping", IEEE , 9 July 2020, <<https://1.ieee802.org/tsn/802-1qcr/>>.

Authors' Addresses

Toerless Eckert
Futurewei Technologies USA
2220 Central Expressway
Santa Clara, CA 95050
United States of America

Email: tte@cs.fau.de

Stewart Bryant
University of Surrey ICS

Email: s.bryant@surrey.ac.uk

Andrew G. Malis
Malis Consulting

Email: agmalis@gmail.com

DetNet Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 September 2022

G. Mirsky
Ericsson
M. Chen
Huawei
B. Varga
J. Farkas
Ericsson
7 March 2022

Operations, Administration and Maintenance (OAM) for Deterministic
Networks (DetNet) with MPLS Data Plane
draft-ietf-detnet-mpls-oam-07

Abstract

This document defines format and use principals of the Deterministic Network (DetNet) service Associated Channel (ACH) over a DetNet network with the MPLS data plane. The DetNet service ACH can be used to carry test packets of active Operations, Administration, and Maintenance protocols that are used to detect DetNet failures and measure performance metrics.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 8 September 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document.

Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
2.1. Terminology and Acronyms	3
2.2. Keywords	4
3. Active OAM for DetNet Networks with MPLS Data Plane	4
3.1. DetNet Active OAM Encapsulation	5
3.2. DetNet Packet Replication, Elimination, and Ordering Functions Interaction with Active OAM	8
4. Use of Hybrid OAM in DetNet	8
5. OAM Interworking Models	8
5.1. OAM of DetNet MPLS Interworking with OAM of TSN	8
5.2. OAM of DetNet MPLS Interworking with OAM of DetNet IP	9
6. IANA Considerations	10
6.1. DetNet MPLS OAM Flags Registry	10
7. Security Considerations	10
8. Acknowledgment	10
9. References	10
9.1. Normative References	10
9.2. Informational References	11
Authors' Addresses	13

1. Introduction

[RFC8655] introduces and explains Deterministic Networks (DetNet) architecture and how the Packet Replication, Elimination, and Ordering functions (PREOF) can be used to ensure low packet drop ratio in DetNet domain.

Operations, Administration and Maintenance (OAM) protocols are used to detect, localize defects in the network, and monitor network performance. Some OAM functions, e.g., failure detection, work in the network proactively, while others, e.g., defect localization, usually performed on-demand. These tasks achieved by a combination of active and hybrid, as defined in [RFC7799], OAM methods.

Also, this document defines format and use principals of the DetNet service Associated Channel over a DetNet network with the MPLS data plane [RFC8964].

2. Conventions used in this document

2.1. Terminology and Acronyms

The term "DetNet OAM" used in this document interchangeably with longer version "set of OAM protocols, methods and tools for Deterministic Networks".

CW Control Word

DetNet Deterministic Networks

d-ACH DetNet Associated Channel Header

d-CW DetNet Control Word

DNH DetNet Header

GAL Generic Associated Channel Label

G-ACh Generic Associated Channel

OAM: Operations, Administration and Maintenance

PREOF Packet Replication, Elimination, and Ordering Functions

PW Pseudowire

RDI Remote Defect Indication

E2E End-to-end

CFM Connectivity Fault Management

BFD Bidirectional Forwarding Detection

TSN Time-Sensitive Network

F-Label A Detnet "forwarding" label that identifies the LSP used to forward a DetNet flow across an MPLS PSN, e.g., a hop-by-hop label used between label switching routers (LSR).

S-Label A DetNet "service" label that is used between DetNet nodes that implement also the DetNet service sub-layer functions. An S-Label is also used to identify a DetNet flow at DetNet service sub-layer.

Underlay Network or Underlay Layer: The network that provides connectivity between the DetNet nodes. MPLS network providing LSP connectivity between DetNet nodes is an example of the underlay layer.

DetNet Node - a node that is an actor in the DetNet domain. DetNet domain edge node and node that performs PREOF within the domain are examples of DetNet node.

2.2. Keywords

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Active OAM for DetNet Networks with MPLS Data Plane

OAM protocols and mechanisms act within the data plane of the particular networking layer. And thus it is critical that the data plane encapsulation supports OAM mechanisms in such a way to comply with the OAM requirements listed in [I-D.tpmb-detnet-oam-framework]. One of such examples that require special consideration is requirement #5:

DetNet OAM packets MUST be in-band, i.e., follow precisely the same path as DetNet data plane traffic both for unidirectional and bi-directional DetNet paths.

The Det Net data plane encapsulation in transport network with MPLS encapsulation specified in [RFC8964]. For the MPLS underlay network, DetNet flows to be encapsulated analogous to pseudowires (PW) over MPLS packet switched network, as described in [RFC3985], [RFC4385]. Generic PW MPLS Control Word (CW), defined in [RFC4385], for DetNet displayed in Figure 1.

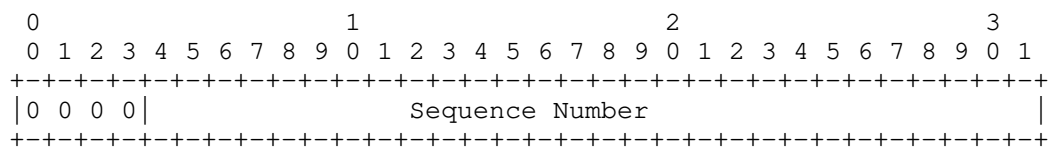


Figure 1: DetNet Control Word Format

PREOF in the DetNet domain composed by a combination of nodes that perform replication and elimination functions. The elimination function always uses the S-Label and packet sequencing information, e.g., the value in the Sequence Number field of DetNet CW (d-CW). The replication sub-function uses the S-Label information only. For data packets Figure 2 presents an example of PREOF in DetNet domain.

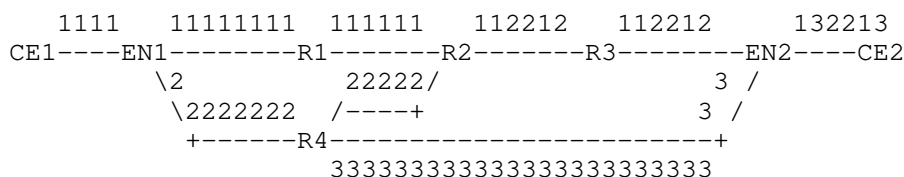


Figure 2: DetNet Data Plane Based on PW

3.1. DetNet Active OAM Encapsulation

DetNet OAM, like PW OAM, uses PW Associated Channel Header defined in [RFC4385]. Figure 3 displays the encapsulation of a DetNet MPLS [RFC8964] active OAM packet.

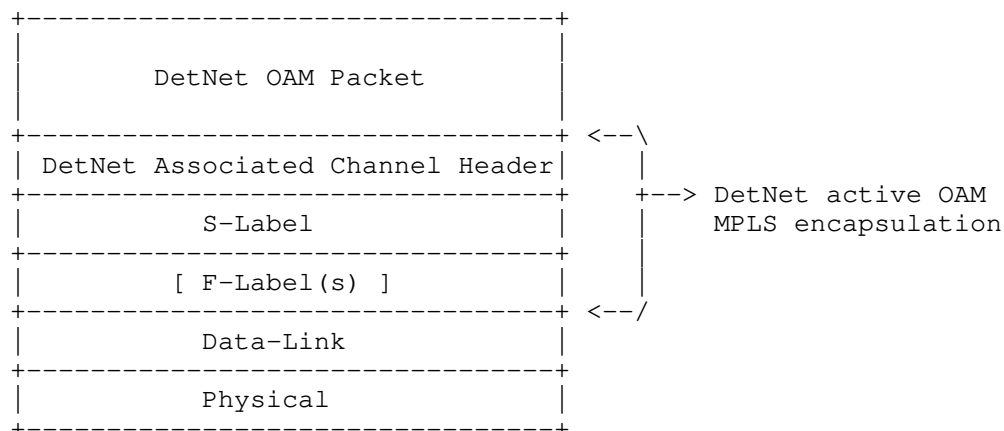


Figure 3: DetNet Active OAM Packet Encapsulation in MPLS Data Plane

Figure 4 displays encapsulation of a test packet of an active DetNet OAM protocol in case of MPLS-over-UDP/IP [RFC9025].

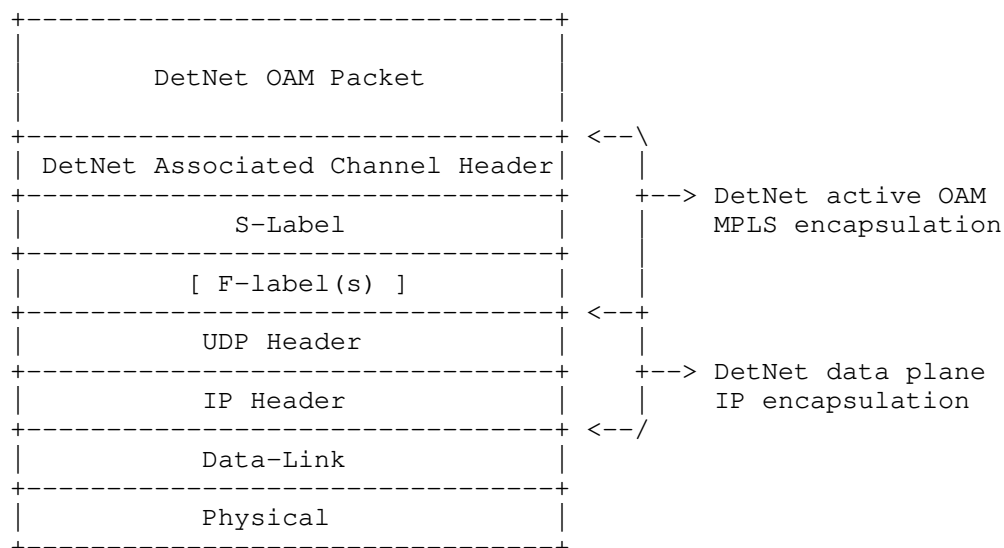


Figure 4: DetNet Active OAM Packet Encapsulation in MPLS-over-UDP/IP

Figure 5 displays the format of the DetNet Associated Channel Header (d-ACH).

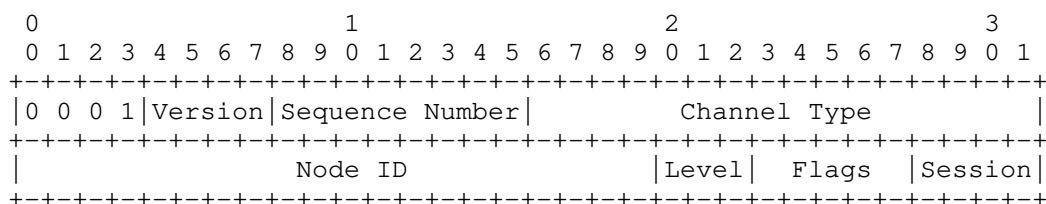


Figure 5: DetNet Associated Channel Header Format

The d-ACH encodes the following fields:

Bits 0..3 MUST be 0b0001. This value of the first nibble allows the packet to be distinguished from an IP packet [RFC4928] and a DetNet data packet [RFC8964].

Version - is a four-bits field, and the value is the version number of the d-ACH. This specification defines version 0x1.

Sequence Number - is an unsigned eight-bit field. The sequence number space is circular with no restriction on the initial value. The originator DetNet node MUST set the value of the Sequence Number field before the transmission of a packet. The originator node MUST increase the value of the Sequence Number field by 1 for each active OAM packet.

Channel Type - contains the value of DetNet Associated Channel Type. It is one of the values defined in the IANA PW Associated Channel Type registry.

Node ID - is an unsigned 20 bits-long field. The value of the Node ID field identifies the DetNet node that originated the packet. Methods of distributing Node ID are outside the scope of this specification.

Level - is a three-bits field.

Flags - is a five-bits field. Flags field contains five one-bit flags. Section 6.1 creates an IANA registry for new flags to be defined. Flags defined in this specification presented in Figure 6.

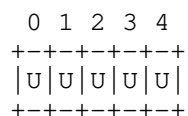


Figure 6: DetNet Associated Channel Header Flags Field Format

U: Unused and for future use. MUST be 0 on transmission and ignored on receipt.

Session ID is a four-bits field.

The DetNet flow, according to [RFC8964], is identified by the S-label that MUST be at the bottom of the stack. Active OAM packet MUST include d-ACH immediately following the S-label.

3.2. DetNet Packet Replication, Elimination, and Ordering Functions Interaction with Active OAM

At the DetNet service sub-layer, special functions MAY be applied to the particular DetNet flow, PREOF, to potentially lower packet loss, improve the probability of on-time packet delivery and ensure in-order packet delivery. PREOF rely on sequencing information in the DetNet service sub-layer. For a DetNet active OAM packet, 28 MSBs of the d-ACH MUST be used as the source of the sequencing information by PREOF.

4. Use of Hybrid OAM in DetNet

Hybrid OAM methods are used in performance monitoring and defined in [RFC7799] as:

Hybrid Methods are Methods of Measurement that use a combination of Active Methods and Passive Methods.

A hybrid measurement method may produce metrics as close to passive, but it still alters something in a data packet even if that is the value of a designated field in the packet encapsulation. One example of such a hybrid measurement method is the Alternate Marking method described in [RFC8321]. Reserving the field for the Alternate Marking method in the DetNet Header will enhance available to an operator set of DetNet OAM tools.

5. OAM Interworking Models

Interworking of two OAM domains that utilize different networking technology can be realized either by a peering or a tunneling model. In a peering model, OAM domains are within the corresponding network domain. When using the peering model, state changes that are detected by a Fault Management OAM protocol can be mapped from one OAM domain into another or a notification, e.g., an alarm, can be sent to a central controller. In the tunneling model of OAM interworking, usually, only one active OAM protocol is used. Its test packets are tunneled through another domain along with the data flow, thus ensuring the fate sharing among test and data packets.

5.1. OAM of DetNet MPLS Interworking with OAM of TSN

Active DetNet OAM is required to provide the E2E fault management and performance monitoring for a DetNet flow. Interworking of DetNet active OAM with MPLS data plane with the IEEE 802.1 Time-Sensitive Networking (TSN) domain based on [RFC9037].

In the case of the peering model is used in the fault management OAM, then the node that borders both TSN and DetNet MPLS domains MUST support [RFC7023]. [RFC7023] specified the mapping of defect states between Ethernet Attachment Circuits (ACs) and associated Ethernet PWs that are part of an end-to-end (E2E) emulated Ethernet service. Requirements and mechanisms described in [RFC7023] are equally applicable to using the peering model to achieve E2E FM OAM over DetNet MPLS and TSN domains. The Connectivity Fault Management (CFM) protocol [IEEE.CFM] or in [ITU.Y1731] can provide fast detection of a failure in the TSN segment of the DetNet service. In the DetNet MPLS domain BFD (Bidirectional Forwarding Detection), specified in [RFC5880] and [RFC5885], can be used. To provide E2E failure detection, the TSN segment might be presented as a concatenated with the DetNet MPLS and the Section 6.8.17 [RFC5880] MAY be used to inform the upstream DetNet MPLS node of a failure of the TSN segment. Performance monitoring can be supported by [RFC6374] in the DetNet MPLS and [ITU.Y1731] in the TSN domains, respectively. Performance objectives for each domain should refer to metrics that additive or be defined for each domain separately.

The following considerations are to be realized when using the tunneling model of OAM interworking between DetNet MPLS and TSN domains:

- * Active OAM test packet MUST be mapped to the same TSN Stream ID as the monitored DetNet flow.
- * Active OAM test packets MUST be treated in the TSN domain based on its S-label and CoS marking (TC field value).

Note that the tunneling model of the OAM interworking requires that the remote peer of the E2E OAM domain supports the active OAM protocol selected on the ingress endpoint. For example, if BFD is used for proactive path continuity monitoring in the DetNet MPLS domain, a TSN endpoint of the DetNet service has also support BFD as defined in [RFC5885].

5.2. OAM of DetNet MPLS Interworking with OAM of DetNet IP

Interworking between active OAM segments in DetNet MPLS and DetNet IP domains can also be realized using either the peering or the tunneling model, as discussed in Section 5.1. Using the same protocol, e.g., BFD, over both segments, simplifies the mapping of errors in the peering model. To provide the performance monitoring over a DetNet IP domain STAMP [RFC8762] and its extensions [RFC8972] can be used.

6. IANA Considerations

6.1. DetNet MPLS OAM Flags Registry

This document describes a new IANA-managed registry to identify DetNet MPLS OAM Flags Bits. The registration procedure is "IETF Review" [RFC8126]. The registry name is "DetNet MPLS OAM Flags". There are five flags in the five-bit Flags field, defined as in Table 1.

Bit	Description	Reference
0-4	Unassigned	This document

Table 1: DetNet MPLS OAM Flags

7. Security Considerations

Additionally, security considerations discussed in DetNet specifications: [RFC8655], [RFC9055], [RFC8964] are applicable to this document. Security concerns and issues related to MPLS OAM tools like LSP Ping [RFC8029], BFD over PW [RFC5885] also apply to this specification.

8. Acknowledgment

Authors extend their appreciation to Pascal Thubert for his insightful comments and productive discussion that helped to improve the document.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC7023] Mohan, D., Ed., Bitar, N., Ed., Sajassi, A., Ed., DeLord, S., Niger, P., and R. Qiu, "MPLS and Ethernet Operations, Administration, and Maintenance (OAM) Interworking", RFC 7023, DOI 10.17487/RFC7023, October 2013, <<https://www.rfc-editor.org/info/rfc7023>>.

- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.
- [RFC8964] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., Bryant, S., and J. Korhonen, "Deterministic Networking (DetNet) Data Plane: MPLS", RFC 8964, DOI 10.17487/RFC8964, January 2021, <<https://www.rfc-editor.org/info/rfc8964>>.
- [RFC9025] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., and S. Bryant, "Deterministic Networking (DetNet) Data Plane: MPLS over UDP/IP", RFC 9025, DOI 10.17487/RFC9025, April 2021, <<https://www.rfc-editor.org/info/rfc9025>>.

9.2. Informational References

- [I-D.tpmb-detnet-oam-framework] Mirsky, G., Theoleyre, F., Papadopoulos, G. Z., and C. J. Bernardos, "Framework of Operations, Administration and Maintenance (OAM) for Deterministic Networking (DetNet)", Work in Progress, Internet-Draft, draft-tpmb-detnet-oam-framework-01, 30 March 2021, <<https://datatracker.ietf.org/doc/html/draft-tpmb-detnet-oam-framework-01>>.
- [IEEE.CFM] IEEE, "Connectivity Fault Management clause of IEEE 802.1Q", IEEE 802.1Q, 2013.
- [ITU.Y1731] ITU-T, "OAM functions and mechanisms for Ethernet based Networks", ITU-T Recommendation G.8013/Y.1731, November 2013.
- [RFC3985] Bryant, S., Ed. and P. Pate, Ed., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, DOI 10.17487/RFC3985, March 2005, <<https://www.rfc-editor.org/info/rfc3985>>.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, DOI 10.17487/RFC4385, February 2006, <<https://www.rfc-editor.org/info/rfc4385>>.

- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, DOI 10.17487/RFC4928, June 2007, <<https://www.rfc-editor.org/info/rfc4928>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC5885] Nadeau, T., Ed. and C. Pignataro, Ed., "Bidirectional Forwarding Detection (BFD) for the Pseudowire Virtual Circuit Connectivity Verification (VCCV)", RFC 5885, DOI 10.17487/RFC5885, June 2010, <<https://www.rfc-editor.org/info/rfc5885>>.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <<https://www.rfc-editor.org/info/rfc6374>>.
- [RFC7799] Morton, A., "Active and Passive Metrics and Methods (with Hybrid Types In-Between)", RFC 7799, DOI 10.17487/RFC7799, May 2016, <<https://www.rfc-editor.org/info/rfc7799>>.
- [RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <<https://www.rfc-editor.org/info/rfc8029>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.
- [RFC8321] Fioccola, G., Ed., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8321, DOI 10.17487/RFC8321, January 2018, <<https://www.rfc-editor.org/info/rfc8321>>.
- [RFC8762] Mirsky, G., Jun, G., Nydell, H., and R. Foote, "Simple Two-Way Active Measurement Protocol", RFC 8762, DOI 10.17487/RFC8762, March 2020, <<https://www.rfc-editor.org/info/rfc8762>>.

- [RFC8972] Mirsky, G., Min, X., Nydell, H., Foote, R., Masputra, A., and E. Ruffini, "Simple Two-Way Active Measurement Protocol Optional Extensions", RFC 8972, DOI 10.17487/RFC8972, January 2021, <<https://www.rfc-editor.org/info/rfc8972>>.
- [RFC9037] Varga, B., Ed., Farkas, J., Malis, A., and S. Bryant, "Deterministic Networking (DetNet) Data Plane: MPLS over IEEE 802.1 Time-Sensitive Networking (TSN)", RFC 9037, DOI 10.17487/RFC9037, June 2021, <<https://www.rfc-editor.org/info/rfc9037>>.
- [RFC9055] Grossman, E., Ed., Mizrahi, T., and A. Hacker, "Deterministic Networking (DetNet) Security Considerations", RFC 9055, DOI 10.17487/RFC9055, June 2021, <<https://www.rfc-editor.org/info/rfc9055>>.

Authors' Addresses

Greg Mirsky
Ericsson
Email: gregimirsky@gmail.com

Mach(Guoyi) Chen
Huawei
Email: mach.chen@huawei.com

Balazs Varga
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary
Email: balazs.a.varga@ericsson.com

Janos Farkas
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary
Email: janos.farkas@ericsson.com

DetNet
Internet-Draft
Intended status: Informational
Expires: 17 April 2022

G. Mirsky
Ericsson
F. Theoleyre
CNRS
G.Z. Papadopoulos
IMT Atlantique
CJ. Bernardos
UC3M
B. Varga
J. Farkas
Ericsson
14 October 2021

Framework of Operations, Administration and Maintenance (OAM) for
Deterministic Networking (DetNet)
draft-ietf-detnet-oam-framework-05

Abstract

Deterministic Networking (DetNet), as defined in RFC 8655, is aimed to provide a bounded end-to-end latency on top of the network infrastructure, comprising both Layer 2 bridged and Layer 3 routed segments. This document's primary purpose is to detail the specific requirements of the Operation, Administration, and Maintenance (OAM) recommended to maintain a deterministic network. With the implementation of the OAM framework in DetNet, an operator will have a real-time view of the network infrastructure regarding the network's ability to respect the Service Level Objective, such as packet delay, delay variation, and packet loss ratio, assigned to each DetNet flow.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 17 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Terminology	3
1.2. Acronyms	4
1.3. Requirements Language	5
2. Role of OAM in DetNet	5
3. Operation	6
3.1. Information Collection	7
3.2. Continuity Check	7
3.3. Connectivity Verification	7
3.4. Route Tracing	8
3.5. Fault Verification/detection	8
3.6. Fault Localization and Characterization	8
3.7. Use of Hybrid OAM in DetNet	9
4. Administration	9
4.1. Collection of metrics	10
4.2. Worst-case metrics	10
5. Maintenance	10
5.1. Replication / Elimination	10
5.2. Resource Reservation	11
5.3. Soft transition after reconfiguration	11
6. Requirements	11
6.1. Requirements on OAM for DetNet Forwarding Sub-layer	12
6.2. Requirements on OAM for DetNet Service Sub-layer	12
7. IANA Considerations	13
8. Security Considerations	13
9. Acknowledgments	13
10. References	13
10.1. Normative References	13
10.2. Informative References	14
Authors' Addresses	15

1. Introduction

Deterministic Networking (DetNet) [RFC8655] has proposed to provide a bounded end-to-end latency on top of the network infrastructure, comprising both Layer 2 bridged and Layer 3 routed segments. That work encompasses the data plane, OAM, time synchronization, management, control, and security aspects.

Operations, Administration, and Maintenance (OAM) Tools are of primary importance for IP networks [RFC7276]. DetNet OAM should provide a toolset for fault detection, localization, and performance measurement.

This document's primary purpose is to detail the specific requirements of the OAM features recommended to maintain a deterministic/reliable network. Specifically, it investigates the requirements for a deterministic network, supporting critical flows.

In this document, the term OAM will be used according to its definition specified in [RFC6291]. DetNet expects to implement an OAM framework to maintain a real-time view of the network infrastructure, and its ability to respect the Service Level Objectives (SLO), such as in-order packet delivery, packet delay, delay variation, and packet loss ratio, assigned to each DetNet flow.

This document lists the functional requirements toward OAM for DetNet domain. The list can further be used for gap analysis of available OAM tools to identify possible enhancements of existing or whether new OAM tools are required to support proactive and on-demand path monitoring and service validation.

1.1. Terminology

This document uses definitions, particularly of a DetNet flow, provided in Section 2.1 [RFC8655]. The following terms are used throughout this document as defined below:

- * DetNet OAM domain: a DetNet network used by the monitored DetNet flow. A DetNet OAM domain (also referred to in this document as "OAM domain") may have MEPs on its edge and MIPs within.
- * DetNet OAM instance: a function that monitors a DetNet flow for defects and/or measures its performance metrics. Within this document, a shorter version, OAM instance, is used interchangeably.

- * Maintenance End Point (MEP): an OAM instance that is capable of generating OAM test packets in the particular sub-layer of the DetNet OAM domain.
- * Maintenance Intermediate endPoint (MIP): an OAM instance along the DetNet flow in the particular sub-layer of the DetNet OAM domain. A MIP MAY respond to an OAM message generated by the MEP at its sub-layer of the same DetNet OAM domain.
- * Control and management plane: the control and management planes are used to configure and control the network (long-term). Relative to a DetNet flow, the control and/or management plane can be out-of-band.
- * Active measurement methods (as defined in [RFC7799]) modify a DetNet flow by inserting novel fields, injecting specially constructed test packets [RFC2544]).
- * Passive measurement methods [RFC7799] infer information by observing unmodified existing flows.
- * Hybrid measurement methods [RFC7799] is the combination of elements of both active and passive measurement methods.
- * In-band OAM is an active OAM is considered in-band in the monitored DetNet OAM domain when it traverses the same set of links and interfaces receiving the same QoS and Packet Replication, Elimination, and Ordering Functions (PREOF) treatment as the monitored DetNet flow.
- * Out-of-band OAM is an active OAM whose path through the DetNet domain is not topologically identical to the path of the monitored DetNet flow, or its test packets receive different QoS and/or PREOF treatment, or both.
- * On-path telemetry can be realized as a hybrid OAM method. The origination of the telemetry information is inherently in-band as packets in a DetNet flow are used as triggers. Collection of the on-path telemetry information can be performed using in-band or out-of-band OAM methods.

1.2. Acronyms

OAM: Operations, Administration, and Maintenance

DetNet: Deterministic Networking

PREOF: Packet Replication, Elimination and Ordering Functions

SLO: Service Level Objective

1.3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Role of OAM in DetNet

DetNet networks expect to provide communications with predictable low packet delay and packet loss. Most critical applications will define an SLO to be required for the DetNet flows it generates.

To respect strict guarantees, DetNet can use an orchestrator able to monitor and maintain the network. Typically, a Software-Defined Network (SDN) controller places DetNet flows in the deployed network based on their SLO. Thus, resources have to be provisioned a priori for the regular operation of the network. OAM represents the essential elements of the network operation and necessary for OAM resources that need to be accounted for to maintain the network operational.

Many legacy OAM tools can be used in DetNet networks, but they are not able to cover all the aspects of deterministic networking. Fulfilling strict guarantees is essential for DetNet flows, resulting in new DetNet specific functionalities that must be covered with OAM. Filling these gaps is inevitable and needs accurate consideration of DetNet specifics. Similar to DetNet flows itself, their OAM needs careful end-to-end engineering as well.

For example, appropriate placing of MEPs along the path of a DetNet flow is not always a trivial task and may require proper design together with the design of the service component of a given DetNet flow.

There are several DetNet specific challenges for OAM. Bounded network characteristics (e.g., delay, loss) are inseparable service parameters; therefore, PM is a key topic for DetNet. OAM tools are needed to prove the SLO without impacting the DetNet flow characteristics. A further challenge is the strict resource allocation. Resources used by OAM must be considered and allocated to avoid disturbing DetNet flow(s).

The DetNet Working Group has defined two sub-layers:

DetNet service sub-layer, at which a DetNet service (e.g., service protection) is provided.

DetNet forwarding sub-layer, which optionally provides resource allocation for DetNet flows over paths provided by the underlying network.

OAM mechanisms exist for the DetNet forwarding sub-layer, nonetheless, OAM for the service sub-layer requires new OAM procedures. These new OAM functions must allow, for example, to recognize/discover DetNet relay nodes, to get information about their configuration, and to check their operation or status.

DetNet service sub-layer functions using a sequence number. That creates a challenge for inserting OAM packets in the DetNet flow.

Fault tolerance also assumes that multiple paths could be provisioned to maintain an end-to-end circuit by adapting to the existing conditions. The central controller/orchestrator typically controls the PREOF on a node. OAM is expected to support monitoring and troubleshooting PREOF on a particular node and within the domain.

Note that distributed controllers can also control PREOF in those scenarios where DetNet solutions involve more than one single central controller.

DetNet forwarding sub-layer is based on legacy technologies and has a much better coverage regarding OAM. However, the forwarding sub-layer is terminated at DetNet relay nodes, so the end-to-end OAM state of forwarding may be created only based on the status of multiple forwarding sub-layer segments serving a given DetNet flow (e.g., in case of DetNet MPLS, there may be no end-to-end LSP below the DetNet PW).

3. Operation

OAM features will enable DetNet with robust operation both for forwarding and routing purposes.

It is worth noting that the test and data packets MUST follow the same path, i.e., the connectivity verification has to be conducted in-band without impacting the data traffic. Test packets MUST share fate with the monitored data traffic without introducing congestion in normal network conditions.

3.1. Information Collection

Information about the state of the network can be collected using several mechanisms. Some protocols, e.g., Simple Network Management Protocol, send queries. Others, e.g., YANG-based data models, generate notifications based on the publish-subscribe method. In either way, information is collected and sent to the controller.

Also, we can characterize methods of transporting OAM information relative to the path of data. For instance, OAM information may be transported in-band or out-of-band relative to the DetNet flow. In case of the former, the telemetry information uses resources allocated for the monitored DetNet flow. If an in-band method of transporting telemetry is used, the amount of generated information needs to be carefully analyzed, and additional resources must be reserved. [I-D.ietf-ippm-ioam-data] defines the in-band transport mechanism where telemetry information is collected in the data packet on which information is generated. Two tracing methods are described - end-to-end, i.e., from the ingress and egress nodes, and hop-by-hop, i.e., like end-to-end with additional information from transit nodes. [I-D.ietf-ippm-ioam-direct-export] and [I-D.mirsky-ippm-hybrid-two-step] are examples of out-of-band telemetry transport. In the former case, information is transported by each node traversed by the data packet of the monitored DetNet flow in a specially constructed packet. In the latter, information is collected in a sequence of follow-up packets that traverse the same path as the data packet of the monitored DetNet flow. In both methods, transport of the telemetry can avoid using resources allocated for the DetNet domain.

3.2. Continuity Check

Continuity check is used to monitor the continuity of a path, i.e., that there exists a way to deliver the packets between two MEP A and MEP B. The continuity check detects a network failure in one direction, from the MEP transmitting test packets to the remote egress MEP.

3.3. Connectivity Verification

In addition to the Continuity Check, DetNet solutions have to verify the connectivity. This verification considers additional constraints, i.e., the absence of misconnection. The misconnection error state is entered after several consecutive test packets from other DetNet flows are received. The definition of the conditions of entry and exit for misconnection error state is outside the scope of this document.

3.4. Route Tracing

Ping and traceroute are two ubiquitous tools that help localize and characterize a failure in the network. They help to identify a subset of the list of routers in the route. However, to be predictable, resources are reserved per flow in DetNet. Thus, DetNet needs to define route tracing tools able to track the route for a specific flow. Also, tracing can be used for the discovery of the Path Maximum Transmission Unit or location of elements of PREOF for the particular route in the DetNet domain.

DetNet is NOT RECOMMENDED to use multiple paths or links, i.e., Equal-Cost Multipath (ECMP) [RFC8939]. As the result, OAM in ECMP environment is outside the scope of this document.

3.5. Fault Verification/detection

DetNet expects to operate fault-tolerant networks. Thus, mechanisms able to detect faults before they impact the network performance are needed.

The network has to detect when a fault occurred, i.e., the network has deviated from its expected behavior. While the network must report an alarm, the cause may not be identified precisely. For instance, the end-to-end reliability has decreased significantly, or a buffer overflow occurs.

DetNet OAM mechanisms SHOULD allow a fault detection in real time. They MAY, when possible, predict faults based on current network conditions. They MAY also identify and report the cause of the actual/predicted network failure.

3.6. Fault Localization and Characterization

An ability to localize the network defect and provide its characterization are necessary elements of network operation.

Fault localization, a process of deducing the location of a network failure from a set of observed failure indications, might be achieved, for example, by tracing the route of the DetNet flow in which the network failure was detected. Another method of fault localization can correlate reports of failures from a set of interleaving sessions monitoring path continuity.

Fault characterization is a process of identifying the root cause of the problem. For instance, misconfiguration or malfunction of PREOF elements can be the cause of erroneous packet replication or extra packets being flooded in the DetNet domain.

3.7. Use of Hybrid OAM in DetNet

Hybrid OAM methods are used in performance monitoring and defined in [RFC7799] as:

Hybrid Methods are Methods of Measurement that use a combination of Active Methods and Passive Methods.

A hybrid measurement method may produce metrics as close to passive, but it still alters something in a data packet even if that is the value of a designated field in the packet encapsulation. One example of such a hybrid measurement method is the Alternate Marking method (AMM) described in [RFC8321]. As with all on-path telemetry methods, AMM in a DetNet domain with the IP data plane is natively in-band in respect to the monitored DetNet flow. Because the marking is applied to a data flow, measured metrics are directly applicable to the DetNet flow. AMM minimizes the additional load on the DetNet domain by using nodal collection and computation of performance metrics in combination with optionally using out-of-band telemetry collection for further network analysis.

4. Administration

The network SHOULD expose a collection of metrics to support an operator making proper decisions, including:

- * **Queuing Delay:** the time elapsed between a packet enqueued and its transmission to the next hop.
- * **Buffer occupancy:** the number of packets present in the buffer, for each of the existing flows.

The following metrics SHOULD be collected:

- * per a DetNet flow to measure the end-to-end performance for a given flow. Each of the paths has to be isolated in multipath routing strategies.
- * per path to detect misbehaving path when multiple paths are applied.
- * per device to detect misbehaving device, when it relays the packets of several flows.

4.1. Collection of metrics

DetNet OAM SHOULD optimize the number of statistics / measurements to collected, frequency of collecting. Distributed and centralized mechanisms MAY be used in combination. Periodic and event-triggered collection information characterizing the state of a network MAY be used.

4.2. Worst-case metrics

DetNet aims to enable real-time communications on top of a heterogeneous multi-hop architecture. To make correct decisions, the controller needs to know the distribution of packet losses/delays for each flow, and each hop of the paths. In other words, the average end-to-end statistics are not enough. The collected information must be sufficient to allow the controller to predict the worst-case.

5. Maintenance

In the face of events that impact the network operation (e.g., link up/down, device crash/reboot, flows starting and ending), the DetNet Controller need to perform repair and re-optimization actions in order to permanently ensure the SLO of all active flows with minimal waste of resources. The controller MUST be able to continuously retrieve the state of the network, to evaluate conditions and trends about the relevance of a reconfiguration, quantifying:

the cost of the sub-optimality: resources may not be used optimally (e.g., a better path exists).

the reconfiguration cost: the controller needs to trigger some reconfigurations. For this transient period, resources may be twice reserved, and control packets have to be transmitted.

Thus, reconfiguration may only be triggered if the gain is significant.

5.1. Replication / Elimination

When multiple paths are reserved between two MEPs, packet replication may be used to introduce redundancy and alleviate transmission errors and collisions. For instance, in Figure 1, the source device S is transmitting the packet to both parents, devices A and B. Each MEP will decide to trigger the packet replication, elimination or the ordering process when a set of metrics passes a threshold value.

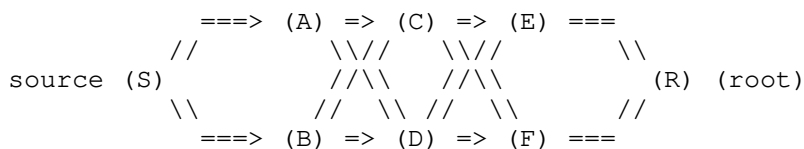


Figure 1: Packet Replication: S transmits twice the same data packet, to DP (A) and AP (B).

5.2. Resource Reservation

Because the quality of service criteria associated with a path may degrade, the network has to provision additional resources along the path. We need to provide mechanisms to patch the network configuration.

5.3. Soft transition after reconfiguration

Since DetNet expects to support real-time flows, DetNet OAM MUST support soft-reconfiguration, where the the additional resources are reserved before the those previously reserved but not in use are released. Some mechanisms have to be proposed so that packets are forwarded through the novel track only when the resources are ready to be used, while maintaining the global state consistent (no packet reordering, duplication, etc.)

6. Requirements

According to [RFC8655], DetNet functionality is divided into forwarding and service sub-layers. The DetNet forwarding sub-layer includes DetNet transit nodes and may allocate resources for a DetNet flow over paths provided by the underlay network. The DetNet service sub-layer includes DetNet relay nodes and provides a DetNet service (e.g., service protection). This section lists general requirements for DetNet OAM as well as requirements in each of the DetNet sub-layers of a DetNet domain.

1. It MUST be possible to initiate a DetNet OAM session from a MEP located at a DetNet node towards downstream MEP(s) within the given domain at a particular DetNet sub-layer. [Ed.note: FT: A MEP may be inside the detnet domain: for instance, for PREOF, an OAM session may be maintained between any pair of replicator / eliminator / egress / ingress.]
2. It MUST be possible to initialize a DetNet OAM session from a centralized controller.

3. DetNet OAM MUST support proactive and on-demand OAM monitoring and measurement methods.
 4. DetNet OAM MUST support unidirectional OAM methods, continuity check, connectivity verification, and performance measurement.
 5. OAM methods MAY combine in-band monitoring or measurement in the forward direction and out-of-bound notification in the reverse direction, i.e., towards the ingress MEP.
 6. DetNet OAM MUST support bi-directional DetNet flows.
 7. DetNet OAM MAY support bi-directional OAM methods for bidirectional DetNet flows. OAM test packets used for monitoring and measurements MUST be in-band in both directions.
 8. DetNet OAM MUST support proactive monitoring of a DetNet device reachability for a given DetNet flow.
 9. DetNet OAM MAY support hybrid performance measurement methods.
 10. DetNet OAM MUST support unidirectional performance measurement methods. Calculated performance metrics MUST include but are not limited to throughput, packet loss, out of order, delay and delay variation metrics. [RFC6374] provides detailed information on performance measurement and performance metrics.
- 6.1. Requirements on OAM for DetNet Forwarding Sub-layer
1. DetNet OAM MUST support Path Maximum Transmission Unit discovery.
 2. DetNet OAM MUST support Remote Defect Indication notification to the DetNet OAM instance performing continuity checking.
 3. DetNet OAM MUST support monitoring levels of resources allocated for the particular DetNet flow. Such resources include but not limited to buffer utilization, scheduler transmission calendar.
 4. DetNet OAM MUST support monitoring any sub-set of paths traversed through the DetNet domain by the DetNet flow.

6.2. Requirements on OAM for DetNet Service Sub-layer

The OAM functions for the DetNet service sub-layer allow, for example, to recognize/discover DetNet relay nodes, to get information about their configuration, and to check their operation or status.

The requirements on OAM for a DetNet relay node are:

1. DetNet OAM MUST provide OAM functions for the DetNet service sub-layer.
 2. DetNet OAM MUST support the discovery of DetNet relay nodes in a DetNet network.
 3. DetNet OAM MUST support the discovery of Packet Replication, Elimination, and Order preservation sub-functions locations in the domain.
 4. DetNet OAM MUST support the collection of the DetNet service sub-layer specific (e.g., configuration/operation/status) information from DetNet relay nodes.
 5. DetNet OAM MUST support exercising functionality of Packet Replication, Elimination, and Order preservation sub-functions in the domain.
 6. DetNet OAM MUST work for DetNet data planes - MPLS and IP.
 7. DetNet OAM MUST support defect notification mechanism, like Alarm Indication Signal. Any DetNet relay node within the given DetNet flow MAY originate a defect notification addressed to any subset of DetNet relay nodes within that flow.
 8. DetNet OAM MUST be able to measure metrics (e.g. delay) inside a collection of OAM sessions, specially for complex DetNet flows, with PREOF features.
7. IANA Considerations
- This document has no actionable requirements for IANA. This section can be removed before the publication.
8. Security Considerations
- This document lists the OAM requirements for a DetNet domain and does not raise any security concerns or issues in addition to ones common to networking and those specific to a DetNet discussed in [RFC9055].
9. Acknowledgments
- The authors express their appreciation and gratitude to Pascal Thubert for the review, insightful questions, and helpful comments.
10. References
- 10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.

10.2. Informative References

- [I-D.ietf-ippm-ioam-data]
Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", Work in Progress, Internet-Draft, draft-ietf-ippm-ioam-data-15, 3 October 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-ippm-ioam-data-15>>.
- [I-D.ietf-ippm-ioam-direct-export]
Song, H., Gafni, B., Zhou, T., Li, Z., Brockners, F., Bhandari, S., Sivakolundu, R., and T. Mizrahi, "In-situ OAM Direct Exporting", Work in Progress, Internet-Draft, draft-ietf-ippm-ioam-direct-export-06, 8 August 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-ippm-ioam-direct-export-06>>.
- [I-D.mirsky-ippm-hybrid-two-step]
Mirsky, G., Lingqiang, W., Zhui, G., and H. Song, "Hybrid Two-Step Performance Measurement Method", Work in Progress, Internet-Draft, draft-mirsky-ippm-hybrid-two-step-11, 8 July 2021, <<https://datatracker.ietf.org/doc/html/draft-mirsky-ippm-hybrid-two-step-11>>.
- [RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, DOI 10.17487/RFC2544, March 1999, <<https://www.rfc-editor.org/info/rfc2544>>.

- [RFC6291] Andersson, L., van Helvoort, H., Bonica, R., Romascanu, D., and S. Mansfield, "Guidelines for the Use of the "OAM" Acronym in the IETF", BCP 161, RFC 6291, DOI 10.17487/RFC6291, June 2011, <<https://www.rfc-editor.org/info/rfc6291>>.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <<https://www.rfc-editor.org/info/rfc6374>>.
- [RFC7276] Mizrahi, T., Sprecher, N., Bellagamba, E., and Y. Weingarten, "An Overview of Operations, Administration, and Maintenance (OAM) Tools", RFC 7276, DOI 10.17487/RFC7276, June 2014, <<https://www.rfc-editor.org/info/rfc7276>>.
- [RFC7799] Morton, A., "Active and Passive Metrics and Methods (with Hybrid Types In-Between)", RFC 7799, DOI 10.17487/RFC7799, May 2016, <<https://www.rfc-editor.org/info/rfc7799>>.
- [RFC8321] Fioccola, G., Ed., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8321, DOI 10.17487/RFC8321, January 2018, <<https://www.rfc-editor.org/info/rfc8321>>.
- [RFC8939] Varga, B., Ed., Farkas, J., Berger, L., Fedyk, D., and S. Bryant, "Deterministic Networking (DetNet) Data Plane: IP", RFC 8939, DOI 10.17487/RFC8939, November 2020, <<https://www.rfc-editor.org/info/rfc8939>>.
- [RFC9055] Grossman, E., Ed., Mizrahi, T., and A. Hacker, "Deterministic Networking (DetNet) Security Considerations", RFC 9055, DOI 10.17487/RFC9055, June 2021, <<https://www.rfc-editor.org/info/rfc9055>>.

Authors' Addresses

Greg Mirsky
Ericsson

Email: gregimirsky@gmail.com

Fabrice Theoleyre
CNRS
300 boulevard Sebastien Brant - CS 10413

67400 Illkirch - Strasbourg
France

Phone: +33 368 85 45 33
Email: theoleyre@unistra.fr
URI: <http://www.theoleyre.eu>

Georgios Z. Papadopoulos
IMT Atlantique
Office B00 - 102A
2 Rue de la Châtaigneraie
35510 Cesson-Sévigné - Rennes
France

Phone: +33 299 12 70 04
Email: georgios.papadopoulos@imt-atlantique.fr

Carlos J. Bernardos
Universidad Carlos III de Madrid
Av. Universidad, 30
28911 Leganes, Madrid
Spain

Phone: +34 91624 6236
Email: cjbc@it.uc3m.es
URI: <http://www.it.uc3m.es/cjbc/>

Balazs Varga
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary

Email: balazs.a.varga@ericsson.com

Janos Farkas
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary

Email: janos.farkas@ericsson.com

Deterministic Networking Working Group
Internet-Draft
Intended status: Informational
Expires: 13 October 2022

P. Liu
China Mobile
Y. Li
Huawei
T. Eckert
Futurewei Technologies USA
Q. Xiong
ZTE Corporation
J. Ryoo
ETRI
11 April 2022

Requirements for Large-Scale Deterministic Networks
draft-liu-detnet-large-scale-requirements-02

Abstract

Aiming at the large-scale deterministic network, this document describes the technical and operational requirements when the different deterministic levels of applications co-exist and are transported over a wide area. This document also describes the corresponding Deterministic Networking (DetNet) data plane enhancement requirements.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 13 October 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	3
2. Conventions Used in This Document	4
3. The Overall Characteristics of Large-Scale Deterministic Networks	4
4. Technical Requirements in Large-Scale Deterministic Networks	6
4.1. Tolerate Time Asynchrony	6
4.1.1. Support Asynchronous Clocks Across Domains	6
4.1.2. Tolerate Clock Jitter & Wander within a Clock Synchronous Domain	7
4.1.3. Provide Mechanisms not Requiring Full Time Synchronization	7
4.1.4. Support Asynchronization based Methods	7
4.2. Support Large Single-hop Propagation Latency	7
4.3. Accommodate the Higher Link Speed	8
4.4. Be Scalable to Numerous Network Devices and Massive Traffic Flows	9
4.5. Tolerate Failures of Links or Nodes and Topology Changes	10
4.6. Support Configuration of Multiple Queueing Mechanisms	10
4.7. Support Queueing Mechanisms Switchover Crossing Multi-domains	11
5. Data Plane Enhancement Requirements	11
5.1. Support Aggregated Flow Identification	11
5.2. Support Queueing Related Information	12
5.3. Support Redundancy Related Fields	12
5.4. Support Explicit Path Selection	12
6. Conclusion	12
7. Security Considerations	12
8. IANA Considerations	13
9. Acknowledgements	13
10. Contributors	13
11. Normative References	13
Appendix A. Examples of Large-Scale Deterministic Network Trials	15
Authors' Addresses	16

1. Introduction

Packet networks are evolving from bandwidth-guaranteed Quality of Service (QoS) to latency-guaranteed QoS that guarantees bounded latency and definite latency. Bounded latency and definite latency can be further understood as in-time delivery, in which a packet arrives without exceeding a predetermined time, and on-time delivery, in which a packet arrives at a predetermined time, respectively. In addition, network survivability, which typically guarantees traffic recovery within 50 ms in the event of a network failure, is evolving to a level that guarantees lossless recovery. In order to realize the evolution of QoS and network survivability of these networks, Time-Sensitive Networking (TSN) technology and Deterministic Networking (DetNet) technology are considered to be essential.

TSN is a set of standards developed by the IEEE 802.1 TSN Task Group (TG) [IEEE802.1TSN] and specifies mechanisms and protocols necessary to realize highly available IEEE 802.1 networks with bounded latency to carry time-sensitive, real-time application traffic.

DetNet, of which architecture is defined in RFC 8655 [RFC8655], provides a capability to carry specified unicast or multicast data flows for real-time applications with extremely low data loss rates and bounded latency within a network domain. The overall framework for DetNet data plane is provided in [RFC8938], and various documents on different data plane technologies and their interworking technologies to extend the service range of data that TSN intends to deliver to the IP (Internet Protocol) and MPLS (Multi-Protocol Label Switching) networks have been standardized.

Since TSN and DetNet were proposed, application use cases have always been one of the hottest topics. As documented in RFC 8578 [RFC8578], the scope of networks addressed by the current DetNet is limited to networks that can be centrally controlled, i.e., an "enterprise" (aka "corporate") network, excluding "the open Internet," explicitly. After years of development, TSN has been used in several industries, and has enough public awareness of the industry for its scope. DetNet also has done a lot of work and the standards are mature, and people become concerned about how to meet deterministic service demand in large-scale networks. The current DetNet is limited to a single administrative domain network, and there are technical elements necessary for application to a large-scale network spanning multiple domains.

This document describes requirements for large-scale deterministic networks where different deterministic levels of applications co-exist and large-scale deterministic networking across multiple administrative domains is possible. This document also describes the requirements for enhancing the DetNet data plane defined prior to this document.

2. Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119][RFC8174] when, and only when, they appear in all capitals, as shown here.

While [RFC2119] and [RFC8174] describe interpretations of these key words in terms of protocol specifications and implementations, they are used in this document to describe technical and operational requirements to realize large-scale deterministic networks.

3. The Overall Characteristics of Large-Scale Deterministic Networks

When deterministic network services are introduced, network providers always face the problem of how to match application needs to the technology, so more works are needed for network service providers to successfully sell DetNet type services to customers. The providers are in need of the following:

Service level objective definitions, considering absolute or relative latency and jitter bounds, flows types and physical network scale

Suitable queuing mechanisms, considering more options for queuing mechanisms for different service level, and

Deployment strategies, considering how to integrate into existing networks, service, and control plane.

[RFC8578] provides various use cases and their requirements in the areas of industry, electricity, buildings, etc. Some of them clearly specify the requirements for latency and jitter, while some others do not for the jitter. Different types of users have different demands, just as a network provider provides different network services for personal business or enterprise business.

One kind has critical SLA requirement, such as remote control or cloud Programmable Logic Controller (PLC) of manufacturing and differential protection of electricity. If these services exceed the boundaries of latency and jitter, it will bring property losses and security risks, so they cannot tolerate with any non-deterministic situation and can pay more on the network service.

Another kind has relatively loose levels of SLA requirement, such as cloud gaming, cloud VR and online meeting for "consumer" networks. The users of these applications hope to have a better network experience, but they can tolerate it to a certain extent. If the network quality is not good sometime, they might be willing to spend more money for high-quality network services. In some aspects, because such services have no industry barriers and can tolerate exceeding the upper boundary of latency within a small probability, they have relatively lower requirements for the network and may be easier to deploy.

Different application demands are actually related to cost. For strict deterministic services, strict technologies need to be used, and all network devices may need to be upgraded. For non-strict deterministic services, it may only be necessary to upgrade some network devices (maybe edge nodes) or share corresponding network resources. From the perspective of deployment, it is helpful if there is a clear classification of application demands, including latency, jitter, reliability, etc. In this way, the corresponding technology to implement could be chosen, taking into account both performance and cost, but how to make choice is not within the scope of this document.

Critical latency requirements:

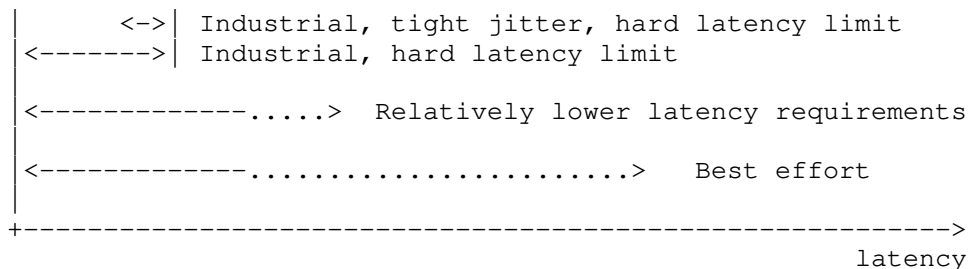


Figure 1: Figure 1: Different levels of application requirements

4. Technical Requirements in Large-Scale Deterministic Networks

Due to the different kinds of application requirements in large-scale networks, the corresponding technical requirements should be considered.

4.1. Tolerate Time Asynchrony

4.1.1. Support Asynchronous Clocks Across Domains

A large-scale network may span over multiple networks with one or more administrative domains. One of DetNet's objectives is to stitch TSN islands together. All devices inside a TSN domain are time-synchronized, and most of TSN technologies rely on precise time synchronization [IEEE802.1Qbv] [IEEE802.1Qch] [IEEE802.1Qav]. However, different TSN islands may have different clocks which are not synchronized as shown in Figure 2, where the time difference of two TSN domains is D . DetNet needs to connect these two TSN domains together and provide end-to-end deterministic latency service. The mechanism adopted by a large-scale deterministic network MUST support the interaction across time domains, so that time domains are synchronized. This can be done, for example, by putting extra buffer space at the ingress of a new domain, increasing the dead time as a guard band, or using some timing compensation mechanism. This document does not intend to list all the potential ways.

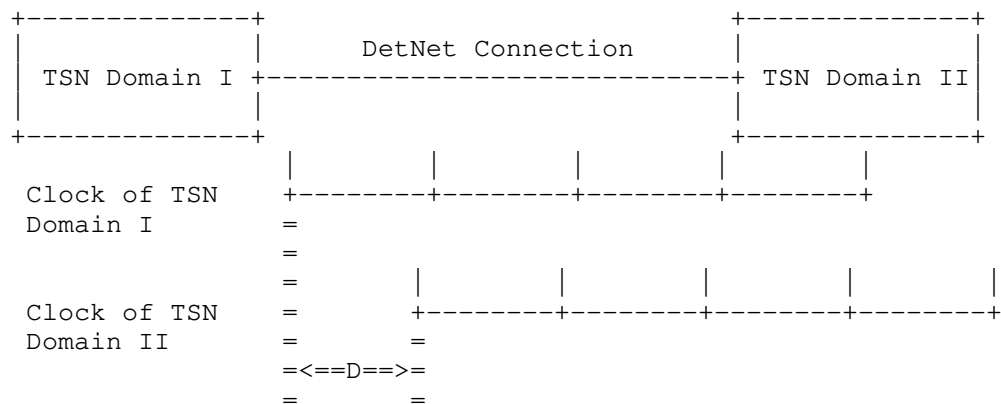


Figure 2: Figure 2: Clock asynchrony between two TSN islands

4.1.2. Tolerate Clock Jitter & Wander within a Clock Synchronous Domain

Within a single time synchronization domain, different clock accuracy is expected, for example the crystal oscillator in Ethernet is specified at 100 ppm[Fast-Ethernet-MII-clock], Synchronous Ethernet (SyncE) can achieve 50 ppb[G.8262], and more precise time synchronization[G.8273] is expected in 5G mobile backhaul. The clocks experience different jitter and wander. It may cause different level of asymmetry of the path. The large-scale networks SHOULD be able to recover or absorb such time variance within a domain and across multiple domains.

4.1.3. Provide Mechanisms not Requiring Full Time Synchronization

Some networks like mobile backhaul use frequency synchronization, such as SyncE, instead of the strict time synchronization. It is usually hard to achieve the full time synchronization in large-scale networks when considering the size of the network topology. It is desired that the same deterministic performance in term of the bounded latency and jitter SHOULD be achieved when full time synchronization is not available, that is to say, when only partial synchronization (SyncE is one of the examples) is in use.

4.1.4. Support Asynchronization based Methods

There are a large number of traffic flows in a large-scale network and some of them are acyclic. Asynchronization based methods can meet the requirements of those traffic flows. Moreover, The mechanisms not requiring the time and/or frequency synchronization eliminate the hardware cost and difficulty at the network nodes.[IEEE802.1Qcr] conceptually uses per-flow based asynchronous shaper to achieve bounded latency. The formula proof shows its effectiveness. It can naturally tolerate the time variance, but it exhibits the concerns of per-flow state buffer management as shown in[I-D.eckert-detnet-bounded-latency-problems]When it is in use, the requirement in Section 4.3 SHOULD be carefully met.

4.2. Support Large Single-hop Propagation Latency

In a large-scale network, a single hop distance is enough to generate large latency. The speed of optical transmission in fiber is 200 km/ms. Thus, the propagation delay of a single hop can be in the order of a few milliseconds. It is much greater than that of a LAN, and introduces impacts on queuing mechanisms, such as cyclic or time aware scheduling method.

For a cyclic based method, suppose a large-scale network wants to keep using the simple cycle mapping relationship, however the link distance between two nodes is longer. Moreover, a downstream node may have many upstream nodes each with different link propagation delays (e.g., 9 us, 10 us, 11 us, 15 us and 20 us). In order to absorb the longest link propagation delay, the length of cycle must be set to at least 20 us. However, since packet's arrival time varies within the receiving cycle, larger cycle length means larger delay variance.

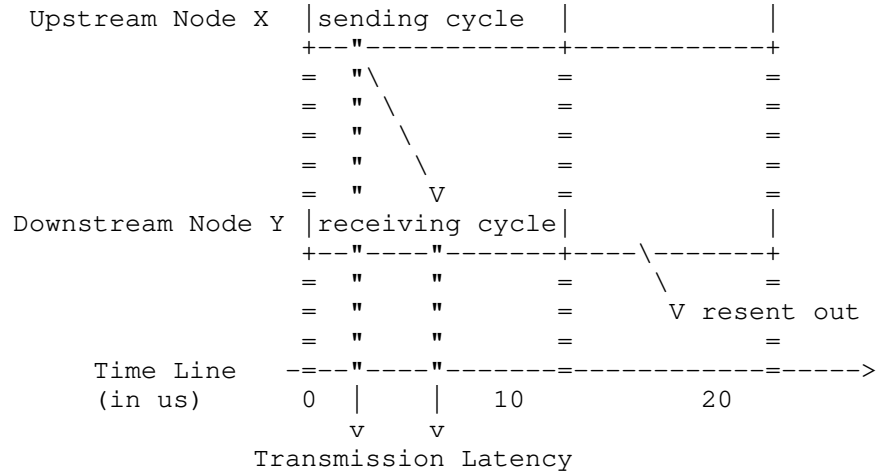


Figure 3: Figure 3: The influence of transmission latency on a cyclic method

4.3. Accommodate the Higher Link Speed

A large-scale network normally uses higher speed links, especially for its backbone. Current deterministic mechanisms used in a local network is usually deployed in link speed of 10 Mbps or 1 Gbps, or possibly 10 Gbps. The data rate of 10G, 100G, 400G and even higher is commonly used in wide area networks. With the increasing of the data rate, the network scheduling cycle can be reduced if the same amount of the data is required to be sent each cycle for each application. Or more data can be sent if the network cycle time remains the same. For the former, it requires the more precise time control (e.g. cycle in the order of a few microseconds or sub-microseconds) for the input stream gate and the timed output buffer. For the latter, more buffer space is required which imposes more complex buffer or queue management and larger memory consumption.

Another aspect to consider is the aggregation of the flows. In the large-scale network, the number of flows can be hundreds or tens of thousands. They can be aggregated into a small number of deterministic path or tunnels. It is practical to have a few flow-based or aggregated-flow based status in the local network. But in higher speed and larger scale networks, it is hardly feasible. If [IEEE802.1Qcr] is in use, it requires more buffers comparing to the other full/partial time synchronized mechanisms. Therefore, it requires optimizations to support higher link speeds.

4.4. Be Scalable to Numerous Network Devices and Massive Traffic Flows

Comparing to a LAN, a large-scale network may have more network devices and traffic flows, and there is a greater possibility of adding or removing network devices and traffic flows. The deterministic latency forwarding mechanisms MUST scale to networks of significant size with numerous network devices and a massive traffic flows.

The increase or decrease of network devices in large-scale networks is more frequent than that in LANs. The change of the number of devices may affect the implementation and adjustment of deterministic network mechanism, such as the topology discovery, queuing mechanism and packet replication and elimination. A simple use case to understand is ultra-low-latency (public) 5G transport networks, which would require DetNet extend to every 5G base station. For some network operators, their networks may need to connect to ~100 K base stations (serving multiple mobile networks operators), and this number will only increase with 5G.

It is almost impossible to identify individual IP flows at the DetNet data plane because of the large overhead and resource reservation for a massive number of flows. DetNet allows the leverage of the flow aggregation. With the large scaling of the network, proper provision at the control plane to accommodate such higher aggregation is required. Individual flows may join and exit the aggregated flow rapidly which causes the dynamic in identification of the aggregated DetNet flow. The wildcards and value ranges used in the identification may have to change in order to ensure the aggregated flows have compatible deterministic characteristics.

The micro-burst will happen more often due to the massive traffic flows, so some methods to decrease it are needed. [I-D.du-detnet-layer3-low-latency] introduces a reference method requiring a scalable buffer to adjust the speed of sending the packets, so as to keep a uniform transmission rate, and it also support the flow aggregation.

4.5. Tolerate Failures of Links or Nodes and Topology Changes

Network link failures are more common in large-scale networks. Path switching or re-convergence of routing will cause high latency of packet loss and retransmission, which is usually in seconds before the network becomes stable again. It is necessary to support certain mechanisms to adapt to failures of links or nodes and topology changes.

The change of path or topology poses a higher challenge to packet replication and elimination. The full disjoint paths when implementing the Packet Replication, Elimination, and Ordering Functions (PREOF) gives a better chance of survival when one of the nodes or links in the path fails. At the same time, it brings the challenges of finding paths with similar distance and/or number of hops so that there is enough buffer space to absorb the latency difference caused by different paths when the scale is large.

4.6. Support Configuration of Multiple Queueing Mechanisms

It is required to provide diversified deterministic service for various applications in a large-scale network and to support the corresponding diversified queueing mechanisms (possibly at multiple DetNet QoS levels). Different queueing mechanisms can provide different levels of latency, jitter and other guarantees, and there may be situations where a network device provides multiple queueing mechanisms at the same time. For example, a network aggregation device may use the mechanisms specified in [IEEE802.1Qbv] and [IEEE802.1Qcr], and other mechanisms to forward traffic to different paths at the same time. By providing a variety of queueing mechanisms to meet diversified deterministic service Requirements, compared with LAN environment, this demand is particularly prominent in large-scale networks. There are usually eight traffic classes in TSN enabled networks. The different queueing mechanisms can be employed to the queues of one or more of those traffic class. In practice, there may be more than eight queues or sub-queues to support more complicated queueing mechanisms.

Accordingly, the configuration for multiple queueing mechanisms is complicated in large-scale deterministic networks and MUST support the unified or simplified scheduling and management of multiple queue mechanisms. For example, in the distributed scenario where there is no controller, flooding the related information of the queue mechanism, including the types and related algorithms, queue forwarding capability, etc. In the centralized scenario, the queueing mechanisms and other information could be reported to the controller to build a deterministic network resource topology pool for path calculation.

4.7. Support Queueing Mechanisms Switchover Crossing Multi-domains

In large-scale deterministic networks, it may across multiple network domains and adopt a variety of different queueing mechanisms within each domain. It is required to support the inter-domain deterministic mechanism at the inter-domain boundary nodes such as the priority redefinition and rescheduling of queues to achieve the end-to-end latency, bounded jitter and packet loss ratio.

Moreover, changing from one queueing mechanism to another may generate additional end-to-end latency and/or jitter which should be taken into consideration. For example, when a flow is forwarded across multiple network domains based on different queueing mechanisms, such as a time synchronous Qbv mechanism [IEEE802.1Qbv] and an asynchronous Qcr mechanism [IEEE802.1Qcr], a collaboration mechanism crossing multi-domains MUST be considered, such as increasing the buffer of inter-domain devices to provide enough adjustment space for the flow to cross different queueing mechanisms, so as to provide end-to-end deterministic services across multiple network domains.

5. Data Plane Enhancement Requirements

According to [RFC8938], the DetNet data plane can provide or carry two metadata in MPLS and IP data planes: Flow-ID and sequence number. The Flow-ID could be used for identification of the DetNet flow or aggregate flow, and the sequence number could be used for PREOF for each DetNet flow. The Flow-ID is used by both the service and forwarding sub-layers, but the sequence number is only used by the service layer. Metadata can also be used for OAM indications and instrumentation of DetNet data plane operation.

Generally speaking, more data plane metadata and related processing SHOULD be supported in the large-scale networks. Native IPv6 data plane should be supported. This section lists the data plane enhancement requirements based on but not limited to the technical requirements in Section 4.

5.1. Support Aggregated Flow Identification

Current IPv6 aggregated flow identification is generally based on 5 or 6 tuples, IP prefixes, or wildcards as indicated in [RFC8938]. However, in large-scale deterministic networks the number of individual flows is huge, and they may randomly join and leave the aggregated flow at each hop. Such behaviours lead to the difficulty in identifying aggregated flows by relying on the prefixes or wildcards.

In addition, flow identification is also used to quickly push a packet to a suitable queue. In a large-scale network, there are mix of flows requiring deterministic latency service and normal forwarding service. Explicit flow identification makes it easier to quickly distinguish the DetNet flows without requiring the longest match rule on multiple tuples in IP data plane. Therefore, explicit aggregated flow identification SHOULD be supported.

5.2. Support Queuing Related Information

According to Section 4.1, a large-scale network should support synchronized or asynchronous queuing mechanisms. Different queueing mechanisms require different metadata to be defined to help regulation and queue management. For instance, the data plane MUST support the identification of cycle for cyclic queuing or the timing related information for time based queuing.

5.3. Support Redundancy Related Fields

Sequence number is the only metadata currently defined for redundancy feature of Detnet. MPLS data plane uses Detnet-over-MPLS label stack to carry it. At the same time, native IPv6 data plane should be able to carry this information too. If specific IP encapsulation or tunnel is in use, this meta data should be defined explicitly for that data plane.

5.4. Support Explicit Path Selection

Explicit route at the control plane and/or management is required so that the "best" path can be selected to meet the latency requirement for DetNet flows. At the data planes, MPLS label stack can be used for this purpose. IP data plane enhancement is required to support the explicit path selection based on IP source routing or SRv6.

6. Conclusion

This document specifies the technical requirements when ensuring the deterministic features in the large-scale networks, and the corresponding data plane enhancement requirements to support the them. Some of the proposed queueing mechanisms and trials are cited and the authors of the document think those proposals give reasonably sound insights to enhancement the current queueing mechanisms to meet the deterministic requirements of the large-scale networks.

7. Security Considerations

There are no IANA actions required by this document.

8. IANA Considerations

This section will be described later.

9. Acknowledgements

The authors would like to thank Yaakov Stein for helpful suggestions. The authors also would like to thank Liang Geng, Peter Willis, Shunsuke Homma and Li Qiang for their previous works.

10. Contributors

The following people have substantially contributed to this document:

Zongpeng Du
China Mobile
EMail: duzongpeng@chinamobile.com

11. Normative References

- [Fast-Ethernet-MII-clock]
"Fast Ethernet MII clock".
- [G.8262] International Telecommunication Union, "Timing characteristics of a synchronous equipment slave clock", ITU-T Recommendation G.8262, November 2018.
- [G.8273] International Telecommunication Union, "Framework of phase and time clocks", ITU-T Recommendation G.8273, March 2018.
- [I-D.dang-queuing-with-multiple-cyclic-buffers]
Liu, B. and J. Dang, "A Queuing Mechanism with Multiple Cyclic Buffers", Work in Progress, Internet-Draft, draft-dang-queuing-with-multiple-cyclic-buffers-00, 22 February 2021, <<https://www.ietf.org/archive/id/draft-dang-queuing-with-multiple-cyclic-buffers-00.txt>>.
- [I-D.du-detnet-layer3-low-latency]
Du, Z. and P. Liu, "Micro-burst Decreasing in Layer3 Network for Low-Latency Traffic", Work in Progress, Internet-Draft, draft-du-detnet-layer3-low-latency-04, 25 October 2021, <<https://www.ietf.org/archive/id/draft-du-detnet-layer3-low-latency-04.txt>>.
- [I-D.eckert-detnet-bounded-latency-problems]
Eckert, T. and S. Bryant, "Problems with existing DetNet bounded latency queuing mechanisms", Work in Progress, Internet-Draft, draft-eckert-detnet-bounded-latency-

problems-00, 12 July 2021,
<<https://www.ietf.org/archive/id/draft-eckert-detnet-bounded-latency-problems-00.txt>>.

[I-D.geng-detnet-requirements-bounded-latency]
Geng, L., Willis, P., Homma, S., and L. Qiang,
"Requirements of Layer 3 Deterministic Latency Service",
Work in Progress, Internet-Draft, draft-geng-detnet-
requirements-bounded-latency-03, 7 July 2019,
<<https://www.ietf.org/archive/id/draft-geng-detnet-requirements-bounded-latency-03.txt>>.

[I-D.qiang-detnet-large-scale-detnet]
Qiang, L., Geng, X., Liu, B., Eckert, T., Geng, L., and G.
Li, "Large-Scale Deterministic IP Network", Work in
Progress, Internet-Draft, draft-qiang-detnet-large-scale-
detnet-05, 2 September 2019,
<<https://www.ietf.org/archive/id/draft-qiang-detnet-large-scale-detnet-05.txt>>.

[IEEE802.1Qav]
IEEE, "IEEE Standard for Local and metropolitan area
networks -- Virtual Bridged Local Area Networks -
Amendment 12: Forwarding and Queuing Enhancements for
Time-Sensitive Streams", IEEE 802.1Qav-2009,
DOI 10.1109/IEEESTD.2010.8684664, 5 January 2010,
<<https://doi.org/10.1109/IEEESTD.2010.8684664>>.

[IEEE802.1Qbv]
IEEE, "IEEE Standard for Local and metropolitan area
networks -- Bridges and Bridged Networks - Amendment 25:
Enhancements for Scheduled Traffic", IEEE 802.1Qbv-2015,
DOI 10.1109/IEEESTD.2016.8613095, 18 March 2016,
<<https://doi.org/10.1109/IEEESTD.2016.8613095>>.

[IEEE802.1Qch]
IEEE, "IEEE Standard for Local and metropolitan area
networks -- Bridges and Bridged Networks - Amendment 29:
Cyclic Queuing and Forwarding", IEEE 802.1Qch-2017,
DOI 10.1109/IEEESTD.2017.7961303, 28 June 2017,
<<https://doi.org/10.1109/IEEESTD.2017.7961303>>.

[IEEE802.1Qcr]
IEEE, "IEEE Standard for Local and Metropolitan Area
Networks -- Bridges and Bridged Networks - Amendment 34:
Asynchronous Traffic Shaping", IEEE 802.1Qcr-2020,
DOI 10.1109/IEEESTD.2020.9253013, 6 November 2020,
<<https://doi.org/10.1109/IEEESTD.2020.9253013>>.

- [IEEE802.1TSN] IEEE Standards Association, "IEEE 802.1 Time-Sensitive Networking Task Group", <<https://www.ieee802.org/1/pages/tsn.html>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8578] Grossman, E., Ed., "Deterministic Networking Use Cases", RFC 8578, DOI 10.17487/RFC8578, May 2019, <<https://www.rfc-editor.org/info/rfc8578>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.
- [RFC8938] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., and S. Bryant, "Deterministic Networking (DetNet) Data Plane Framework", RFC 8938, DOI 10.17487/RFC8938, November 2020, <<https://www.rfc-editor.org/info/rfc8938>>.

Appendix A. Examples of Large-Scale Deterministic Network Trials

Some trials have been carried out to verify the concept of large-scale deterministic networks.

In order to verify the deterministic technology of large-scale networks, a trial of Deterministic IP on China Environment for Network Innovations (CENI), which is a network built for new network technology trial, was deployed. A network with a distance of 3,000 km over 13 hops was tested, and the jitter was controlled within 100us.

In order to verify the remote control on Deterministic IP, which required that the latency should be controlled within 4 ms and jitter should be controlled within 20 us. A trial cooperated with Baosteel spanned 600 km was deployed. Baosteel is a Chinese steel company and put forward this demand. Both of the first and second trials are based on a frequency synchronization solution. The mechanism details could be found in . [I-D.dang-queuing-with-multiple-cyclic-buffers][I-D.qiang-detnet-large-scale-detnet].

In order to realize multi flows synchronization on an inter-provincial network in an exhibition, Emergen proposed the requirement that two flows of video and virtual reality (VR) were sent from province A, and arrived at province B together, so people can see the synchronization of video collected by camera and the VR model. This requirement was proposed to facilitate the virtual industry product deployment. Due to time and other problems, it was realized by the edge network device for a relatively lower levels of service level agreement (SLA).

Teaming up with a smart factory operator, network operators, equipment companies, and universities, ETRI demonstrated an ultra-low latency, high-reliability 5G wired and wireless network-based remote industrial Internet of Things (IIoT) service by connecting a control center and a smart factory through three different operators' networks at a distance of 280 km. In this trail, it was demonstrated that real-time remote smart manufacturing service is possible by making round-trip delay below 3 ms within a smart factory and below 10 ms between remote 5G industrial devices. In the future, the team plans to examine feasibility of large-scale deterministic networking by connecting smart factories in Gyeongsan, South Korea and Oulu, Finland.

These trials show that both operators and enterprise users begin to put forward requirements for the certainty of large-scale networks, but the implementation technologies are not exactly the same.

Authors' Addresses

Peng Liu
China Mobile
Beijing
100053
China
Email: liupengyjy@chinamobile.com

Yizhou Li
Huawei
Nanjing
210012
China
Email: liyizhou@huawei.com

Toerless Eckert
Futurewei Technologies USA
Santa Clara, 95014
United States of America
Email: tte@cs.fau.de

Quan Xiong
ZTE Corporation
Wuhan
430223
China
Email: xiong.quan@zte.com.cn

Jeong-dong Ryoo
ETRI
Daejeon
34129
South Korea
Email: ryoo@etri.re.kr

DetNet
Internet-Draft
Intended status: Informational
Expires: 5 August 2022

B. Varga
J. Farkas
Ericsson
A. Malis
Malis Consulting
1 February 2022

Deterministic Networking (DetNet): DetNet PREOF via MPLS over UDP/IP
draft-varga-detnet-ip-preof-02

Abstract

This document describes how DetNet IP data plane can support the Packet Replication, Elimination, and Ordering Functions (PREOF) built on the existing MPLS PREOF solution [RFC8939] and the mechanisms defined in [RFC9025].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 5 August 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

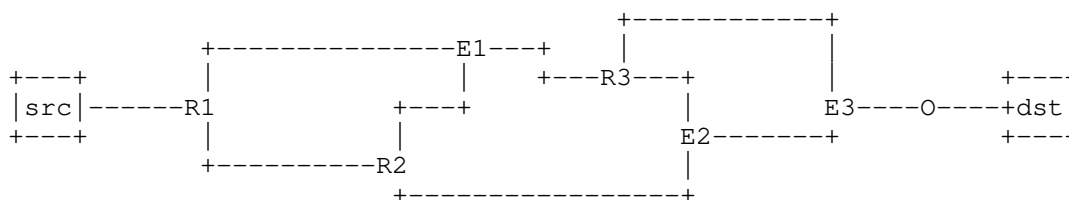
This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
2.1. Terms Used in This Document	3
2.2. Abbreviations	3
2.3. Requirements Language	4
3. Requirements for adding PREOF to DetNet IP	4
4. Adding PREOF to DetNet IP	4
4.1. Solution Basics	4
4.2. Encapsulation	5
4.3. Packet Processing	6
4.4. Flow Aggregation	6
4.5. PREOF Procedures	7
4.6. PREOF capable DetNet IP domain	8
5. Control and Management Plane Parameters	8
6. Security Considerations	10
7. IANA Considerations	10
8. References	10
8.1. Normative References	10
8.2. Informative References	11
Authors' Addresses	11

1. Introduction

The DetNet Working Group has defined packet replication (PRF), packet elimination (PEF) and packet ordering (POF) functions to provide service protection by the DetNet service sub-layer [RFC8655]. The PREOF service protection method relies on copies of the same packet sent over multiple maximally disjoint paths and uses sequencing information to eliminate duplicates. A possible implementation of the PRF and PEF functions is described in [IEEE8021CB] and the related YANG data model is defined in [IEEEP8021CBcv]. A possible implementation of POF function is described in [I-D.varga-detnet-pof]. Figure 1 shows a DetNet flow on which PREOF functions are applied during forwarding from the source to the destination.



R: replication function (PRF)

E: elimination function (PEF)

O: ordering function (POF)

Figure 1: PREOF scenario in a DetNet network

In general, the use of PREOF functions require sequencing information to be included in the packets of a DetNet compound flow. This may be done by adding a sequence number or time stamp as part of DetNet encapsulation. Sequencing information is typically added once, at or close to the source.

The DetNet MPLS data plane [RFC8939] specifies how sequencing information is encoded in the MPLS header. However, the DetNet IP data plane described in [RFC8939] does not specify how sequencing information can be encoded in the IP header. This document describes a DetNet IP encapsulation that includes sequencing information based on the DetNet MPLS over UDP/IP data plane [RFC9025], i.e., leveraging the MPLS-over-UDP technology.

2. Terminology

2.1. Terms Used in This Document

This document uses the terminology established in the DetNet architecture [RFC8655], and the reader is assumed to be familiar with that document and its terminology.

2.2. Abbreviations

The following abbreviations are used in this document:

DetNet Deterministic Networking.

PEF Packet Elimination Function.

POF Packet Ordering Function.

PREOF Packet Replication, Elimination and Ordering Functions.

PRF Packet Replication Function.

2.3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Requirements for adding PREOF to DetNet IP

The requirements for adding PREOF to DetNet IP are:

- * to reuse existing DetNet data plane solutions (e.g., [RFC8964], [RFC9025]).
- * to allow the DetNet service sub-layer for IP packet switched networks with minimal implementation effort.

The described solution practically gains from MPLS header fields without adding MPLS protocol stack complexity to the nodal requirements.

4. Adding PREOF to DetNet IP

4.1. Solution Basics

The DetNet IP encapsulation supporting DetNet Service sub-layer is based on the "UDP tunneling" concept. The solution creates a set of underlay UDP/IP tunnels between an overlay set of DetNet relay nodes.

At the edge of a PREOF capable DetNet IP domain the DetNet flow is encapsulated in an UDP packet containing the sequence number used by PREOF functions within the domain. This solution maintains the 6-tuple-based DetNet flow identification in DetNet transit nodes, which operate at the DetNet forwarding sub-layer between the DetNet service sub-layer nodes; therefore, it is compatible with [RFC8939]. Figure 2 shows how the PREOF capable DetNet IP data plane fits into the DetNet sub-layers.

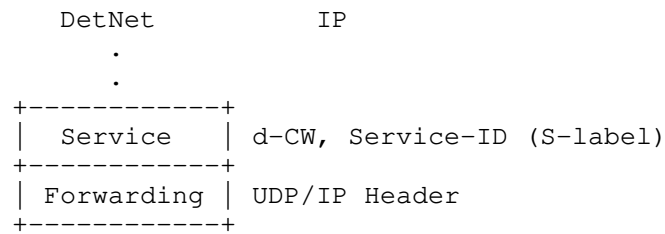


Figure 2: PREOF capable DetNet IP data plane

4.2. Encapsulation

The PREOF capable DetNet IP encapsulation builds on encapsulating DetNet PW directly over UDP. That is, it combines DetNet MPLS [RFC8964] with DetNet MPLS-in-UDP [RFC9025], without using any F-Labels as shown in Figure 3. DetNet flows are identified at the receiving DetNet service sub-layer processing node via the S-Label and/or the UDP/IP header information. Sequencing information for PREOF is provided by the DetNet Control Word (d-CW) as per [RFC8964]. The S-label is used to identify both the DetNet flow and the DetNet App-flow type. The UDP tunnel is used to direct the packet across the DetNet domain to the next DetNet service sub-layer processing node.

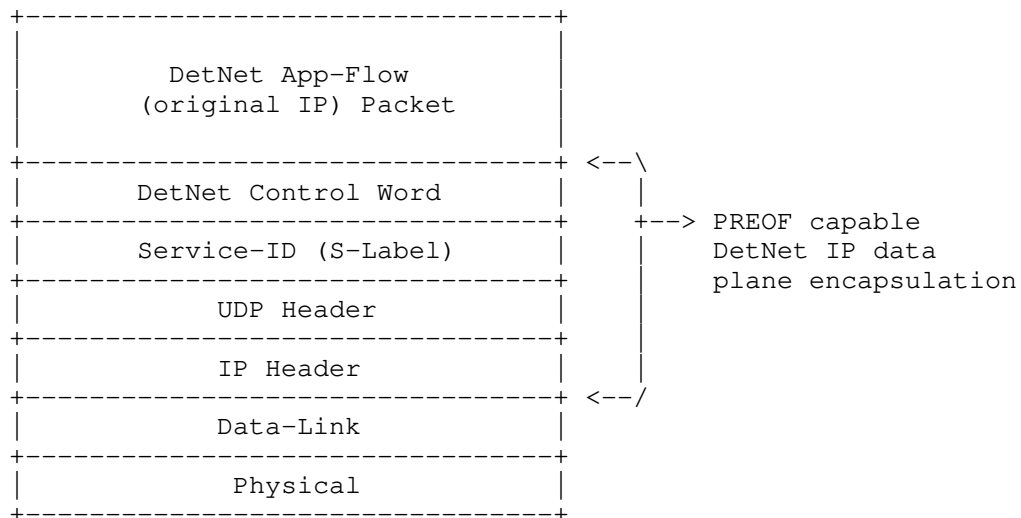


Figure 3: PREOF capable DetNet IP encapsulation

4.3. Packet Processing

IP ingress and egress nodes of the PREOF capable DetNet IP domain MUST add and remove a DetNet service-specific d-CW and Service-ID (i.e., S-Label). Relay nodes MAY change Service-ID values when processing a DetNet flow, i.e., incoming and outgoing Service-IDs of a DetNet flow can be different. Service-ID values MUST be provisioned per DetNet service via configuration, i.e., via the Controller Plane described in [RFC8938]. In some PREOF topologies, the node performing replication sends the packets to multiple nodes performing e.g., PEF or POF and the replication node may need to use different Service-ID values for the different member flows for the same DetNet service.

Note, that Service-IDs provide identification at the downstream DetNet service sub-layer receiver, not the sender.

4.4. Flow Aggregation

Two methods can be used for flow aggregation:

- * aggregation using same UDP tunnel,
- * aggregating DetNet flows as a new DetNet flow.

In the first case, the different DetNet PWs use the same UDP tunnel, so they are treated as a single (aggregated) flow on all transit nodes.

For the second option, an additional Service-ID and d-CW tuple is added to the encapsulation. The Aggregate-ID is a special case of a Service-ID, whose properties are known only at the aggregation and de-aggregation end points. It is a property of the Aggregate-ID that it is followed by a d-CW followed by an Service-ID/d-CW tuple. Figure 4 shows the encapsulation in case of aggregation.

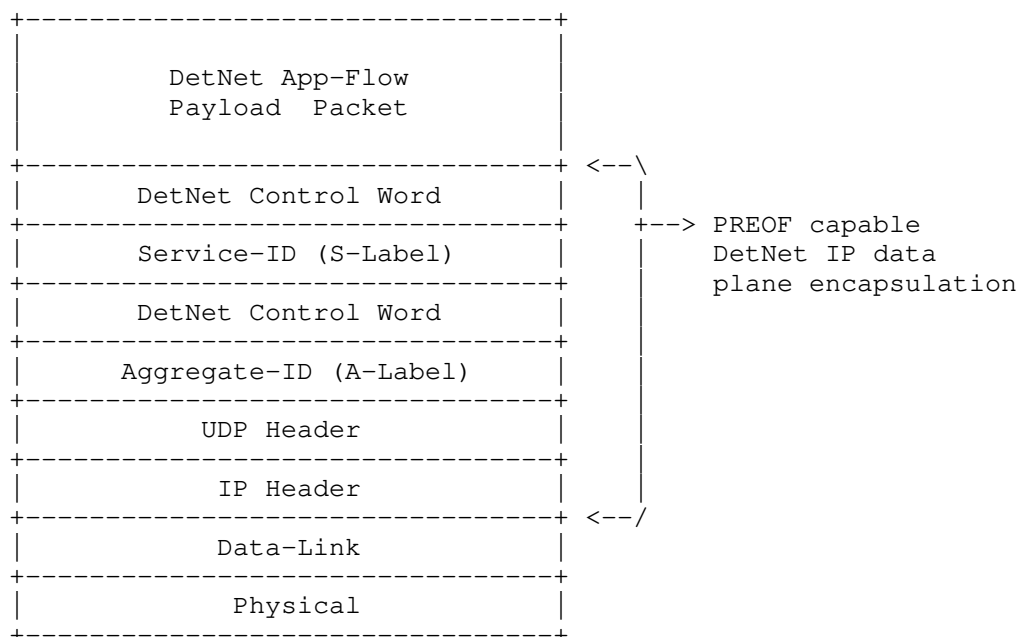


Figure 4: Aggregating DetNet flows as a new DetNet flow

4.5. PREOF Procedures

A node operating on a received DetNet flow at the DetNet service sub-layer uses the local context associated with a received Service-ID to determine which local DetNet operation(s) are applied to received packet. A Service-ID may be allocated to be unique and enabling DetNet flow identification regardless of which input interface or UDP tunnel the packet is received. It is important to note that Service-ID values are driven by the receiver, not the sender.

The DetNet forwarding sub-layer is supported by the UDP tunnel and is responsible for providing resource allocation and explicit routes.

To support outgoing PREOF capable DetNet IP encapsulation, an implementation MUST support the provisioning of UDP and IP header information. Note, when PRF is performed at the DetNet service sub-layer, there are multiple member flows, and each member flow requires their own Service-ID, UDP and IP header information. The headers for each outgoing packet MUST be formatted according to the configuration information, and the UDP Source Port value MUST be set to uniquely identify the DetNet flow. The packet MUST then be handled as a PREOF capable DetNet IP packet.

To support the receive processing, an implementation **MUST** also support the provisioning of received Service-ID, UDP and IP header information. The provisioned information **MUST** be used to identify incoming app-flows based on the combination of Service-ID and/or incoming encapsulation header information.

4.6. PREOF capable DetNet IP domain

Figure 5 shows using PREOF in a PREOF capable DetNet IP network.

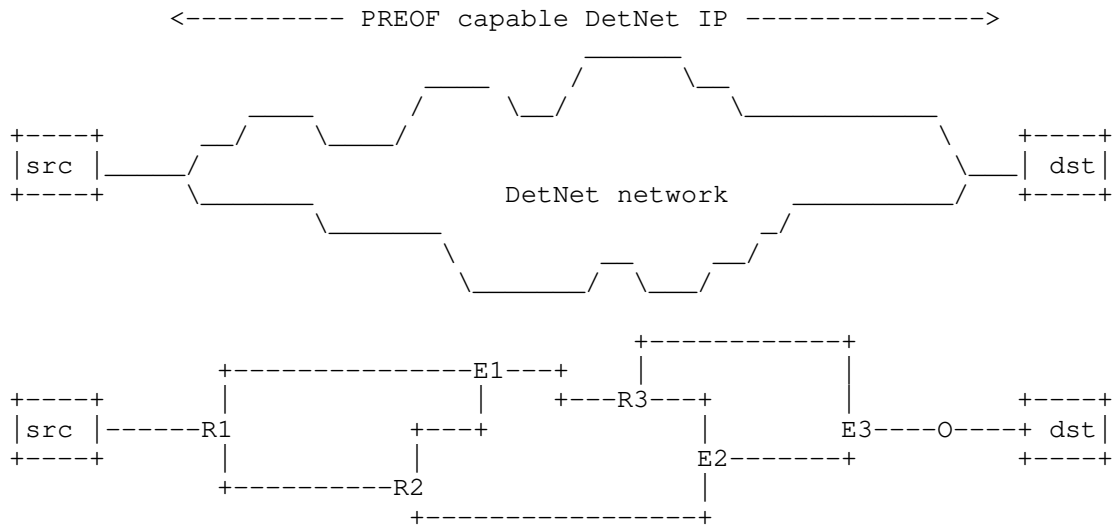


Figure 5: PREOF capable DetNet IP domain

5. Control and Management Plane Parameters

The information needed to identify individual and aggregated DetNet flows is summarized as follows:

- * Service-ID information to be mapped to UDP/IP flows. Note that, for example, a single Service-ID can map to multiple sets of UDP/IP information when PREOF is used.
- * IPv4 or IPv6 source address field.
- * IPv4 or IPv6 source address prefix length, where a zero (0) value effectively means that the address field is ignored.
- * IPv4 or IPv6 destination address field.

- * IPv4 or IPv6 destination address prefix length, where a zero (0) effectively means that the address field is ignored.
- * IPv4 protocol field set to "UDP".
- * IPv6 next header field set to "UDP".
- * For the IPv4 Type of Service and IPv6 Traffic Class Fields:
 - Whether or not the DSCP field is used in flow identification as the use of the DSCP field for flow identification is optional.
 - If the DSCP field is used to identify a flow, then the flow identification information (for that flow) includes a list of DSCPs used by the given DetNet flow.
- * UDP Source Port. Support for both exact and wildcard matching is required. Port ranges can optionally be used.
- * UDP Destination Port. Support for both exact and wildcard matching is required. Port ranges can optionally be used.
- * For end systems, an optional maximum IP packet size that should be used for that outgoing DetNet IP flow.

This information MUST be provisioned per DetNet flow via configuration, e.g., via the controller plane.

An implementation MUST support ordering of the set of information used to identify an individual DetNet flow. This can, for example, be used to provide a DetNet service for a specific UDP flow, with unique Source and Destination Port field values, while providing a different service for the aggregate of all other flows with that same UDP Destination Port value.

The minimum set of information for the configuration of the DetNet service sub-layer is summarized as follows:

- * App-flow identification information.
- * Sequence number length.
- * PREOF + related Service-ID(s).
- * Associated forwarding sub-layer information.
- * Service aggregation information.

The minimum set of information for the configuration of the DetNet forwarding sub-layer is summarized as follows:

- * UDP tunnel specific information.
- * Traffic parameters.

6. Security Considerations

There are no new DetNet related security considerations introduced by this solution.

7. IANA Considerations

This document makes no IANA requests.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.
- [RFC8938] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., and S. Bryant, "Deterministic Networking (DetNet) Data Plane Framework", RFC 8938, DOI 10.17487/RFC8938, November 2020, <<https://www.rfc-editor.org/info/rfc8938>>.
- [RFC8939] Varga, B., Ed., Farkas, J., Berger, L., Fedyk, D., and S. Bryant, "Deterministic Networking (DetNet) Data Plane: IP", RFC 8939, DOI 10.17487/RFC8939, November 2020, <<https://www.rfc-editor.org/info/rfc8939>>.
- [RFC8964] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., Bryant, S., and J. Korhonen, "Deterministic Networking (DetNet) Data Plane: MPLS", RFC 8964, DOI 10.17487/RFC8964, January 2021, <<https://www.rfc-editor.org/info/rfc8964>>.

[RFC9025] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., and S. Bryant, "Deterministic Networking (DetNet) Data Plane: MPLS over UDP/IP", RFC 9025, DOI 10.17487/RFC9025, April 2021, <<https://www.rfc-editor.org/info/rfc9025>>.

8.2. Informative References

[I-D.varga-detnet-pof]
Varga, B., Farkas, J., Kehrer, S., and T. Heer,
"Deterministic Networking (DetNet): Packet Ordering
Function", Work in Progress, Internet-Draft, draft-varga-
detnet-pof-02, 22 October 2021,
<<https://www.ietf.org/archive/id/draft-varga-detnet-pof-02.txt>>.

[IEEE8021CB]
IEEE, "IEEE Standard for Local and metropolitan area
networks -- Frame Replication and Elimination for
Reliability", DOI 10.1109/IEEESTD.2017.8091139, October
2017,
<https://standards.ieee.org/standard/802_1CB-2017.html>.

[IEEEP8021CBcv]
Kehrer, S., "FRER YANG Data Model and Management
Information Base Module", IEEE P802.1CBcv
/D1.2 P802.1CBcv, March 2021,
<<https://www.ieee802.org/1/files/private/cv-drafts/d1/802-1CBcv-d1-2.pdf>>.

Authors' Addresses

Balázs Varga
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary

Email: balazs.a.varga@ericsson.com

János Farkas
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary

Email: janos.farkas@ericsson.com

Andrew G. Malis
Malis Consulting

Email: agmalis@gmail.com

DetNet
Internet-Draft
Intended status: Informational
Expires: 27 October 2022

B. Varga, Ed.
J. Farkas
Ericsson
S. Kehrer
T. Heer
Hirschmann Automation and Control GmbH
25 April 2022

Deterministic Networking (DetNet): Packet Ordering Function
draft-varga-detnet-pof-03

Abstract

Replication and Elimination functions of DetNet [RFC8655] may result in out-of-order packets, which may not be acceptable for some time-sensitive applications. The Packet Ordering Function (POF) algorithm described herein enables to restore the correct packet order when replication and elimination functions are used in DetNet networks.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 27 October 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
2.1. Terms Used in This Document	3
2.2. Abbreviations	4
2.3. Requirements Language	4
3. Requirements on POF Implementations	4
4. POF Algorithms	5
4.1. Prerequisites and Assumptions	5
4.2. POF building blocks	5
4.3. The Basic POF Algorithm	6
4.4. The Advanced POF Algorithm	8
4.5. Further enhancements of POF algorithms	9
4.6. Selecting and using the POF algorithm	9
5. Control and Management Plane Parameters for POF	10
6. Security Considerations	10
7. IANA Considerations	10
8. Acknowledgements	10
9. References	10
9.1. Normative References	10
9.2. Informative References	11
Authors' Addresses	11

1. Introduction

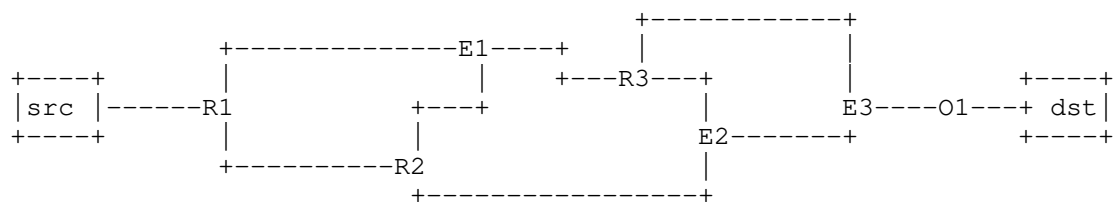
The DetNet Working Group has defined packet replication (PRF) and packet elimination (PEF) functions for achieving extremely low packet loss. PRF and PEF are described in [RFC8655] and provide service protection for DetNet flows. This service protection method relies on copies of the same packet sent over multiple maximally disjoint paths and uses sequencing information to eliminate duplicates. A possible implementation of PRF and PEF functions is described in [IEEE8021CB] and the related YANG model is defined in [IEEEP8021CBcv].

In general, use of per packet replication and elimination functions may result in out-of-order delivery of packets, which may not be acceptable for some deterministic applications. Correcting packet

order is not a trivial task, therefore details of a Packet Ordering Function (POF) are specified herein. The IETF DetNet WG has defined in [RFC8655] the external observable result of a POF function, i.e., that packets are reordered, but without any implementation details.

So far in packet networks, out-of-order delivery situations were handled at higher OSI layers at the end-points/hosts (e.g., in the TCP stack when packets are sent to application layer) and not within a network in nodes acting at the Layer-2 or Layer-3 OSI layers.

Figure 1 shows a DetNet flow on which PREOF functions are applied during forwarding from source to destination.



R: replication point (PRF)

E: elimination point (PEF)

O: ordering function (POF)

Figure 1: PREOF scenario in a DetNet network

Important to note, that application may react differently on out-of-order delivery. A single out-of-order packet (E.g., packet order: #1, #3, #2, #4, #5) may be interpreted by some applications as a single error, but some other applications may treat it as a 3 errors in-a-row situation. 3 errors in-a-row is a usual error threshold and may cause the application to stop (e.g., to transition to a fail safe state).

2. Terminology

2.1. Terms Used in This Document

This document uses the terminology established in the DetNet architecture [RFC8655], and the reader is assumed to be familiar with that document and its terminology.

2.2. Abbreviations

The following abbreviations are used in this document:

DetNet	Deterministic Networking.
PEF	Packet Elimination Function.
POF	Packet Ordering Function.
PREOF	Packet Replication, Elimination and Ordering Functions.
PRF	Packet Replication Function.

2.3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Requirements on POF Implementations

The requirements on a POF function are:

- * to solve the out-of-order delivery problem of the Replication and Elimination functions of DetNet networks.
- * to consider the delay bound requirement of a DetNet Flow.
- * to be simple and to require in network nodes only a minimum set of states/configuration parameters and resources per DetNet Flow.
- * to add only minimal or no delay to the forwarding process of packets.
- * not to require synchronization between PREOF nodes.

Some aspects are explicitly out-of-scope for a POF function:

- * to eliminate the delay variation caused by the packet ordering. Dealing with delay variation is a DetNet forwarding sub-layer target and it can be achieved for example by placing a de-jitter buffer or flow regulator (e.g., shaping) function after the POF functionality.

4. POF Algorithms

4.1. Prerequisites and Assumptions

The POF Algorithm discussed in this document makes some assumptions and tradeoffs regarding the characteristics of the sequence of received packets. In particular, the algorithm assumes that a Packet Elimination Function (PEF) is performed on the incoming packets before they are handed to the POF function. Hence, the sequence of incoming packets can be out of order or incomplete but cannot contain duplicate packets. However, the PREOF functions run independently without any state exchange required between the PEF and the POF or the PRF and the POF. Error cases in which the POF is presented duplicate packets may lead to out of order delivery of duplicate packets as well as to increased delays.

The algorithm further requires that the delay difference between two replicated packets that arrive at the PRF before the POF is bounded and known. Error cases that violate this condition (e.g., a packet that arrives later than this bound) will result in out-of order packets.

The algorithm also makes some tradeoffs. For simplicity, it is designed in a way that allows for some out of order packets directly after initialization. If this is not acceptable, Section 4.5 provides an alternative initialization scheme that prevents out-of-order packets in the initialization phase.

4.2. POF building blocks

The method described herein provides POF for DetNet networks. The configuration parameters of POF can be derived during engineering the DetNet flow through the network.

The POF method is provided via:

1. Conditional buffer: for buffering the out-of-order packets of a DetNet flow for a given time.
2. Delay calculator: buffering time considers (i) the delay difference of paths used for forwarding the replicated packets and (ii) the bounded delay requirement of the given DetNet flow.

Note: the conditional buffer of POF increases the burstiness of the traffic as it adds delay only for some of the packets.

Figure 2 shows the building blocks of a possible POF implementation.

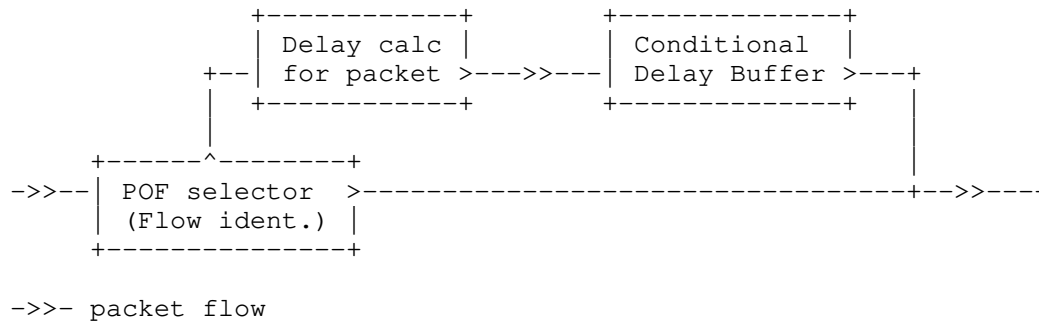


Figure 2: POF Building Blocks

4.3. The Basic POF Algorithm

The basic POF algorithm delays all out-of-order packets until all previous packet arrives or a given time (POFMaxDelay) elapses. The basic POF algorithm works as follows:

- * The sequence number of the last forwarded packet (POFLastSent) is stored for each DetNet Flow.
- * The sequence number (seq_num) of a received packet is compared to that of the last forwarded one (POFLastSent).
- * If (seq_num <= POFLastSent + 1)
 - Then the packet is forwarded and "POFLastSent" is updated (POFLastSent = seq_num).
 - Else the received packet is buffered.
- * A buffered packet is forwarded from the buffer when its seq_num becomes equal to "POFLastSent +1," OR a predefined time ("POFMaxDelay") elapses.
- * When a packet is forwarded from the buffer "POFLastSent" is updated with its seq_num (POFLastSent = seq_num).

Note: the difference of sequence number in consecutive packets is bounded due to the history window of the Elimination function before the POF. Therefore "<=" can be evaluated despite of the circular sequence number space.

The state used by the basic POF algorithm (i.e., "POFLastSent") needs initialization and maintenance. This works as follows:

- * The next received packet must be forwarded and the POFLastSent updated when the POF function was reset OR no packet was received for a predefined time ("POFTakeAnyTime").
- * The reset of POF erases all frames/packets from the time-based buffer used by POF.

The basic POF algorithm has two parameters to engineer:

- * "POFMaxDelay", which cannot be smaller than the delay difference of the paths used by the flow.
- * "POFTakeAnyTime", which is calculated based on several factors, for example the RECOVERY_TIMEOUT related settings of the Elimination function(s) before the POF, the flow characteristics (e.g., inter frame/packet time), and the delay difference of the paths used by the flow.

Design of these parameters is out-of-scope in this document.

Note: multiple network failures may impact the POF function (e.g., complete outage of all redundant paths).

The basic POF algorithm increases the delay of packets with maximum "POFMaxDelay" time. Packets being in order are not delayed. This basic POF method can be applied in all network scenarios where the remaining delay budget of a flow at the POF point is larger than "POFMaxDelay" time.

Figure 3 shows the delay budget relations at the POF point.

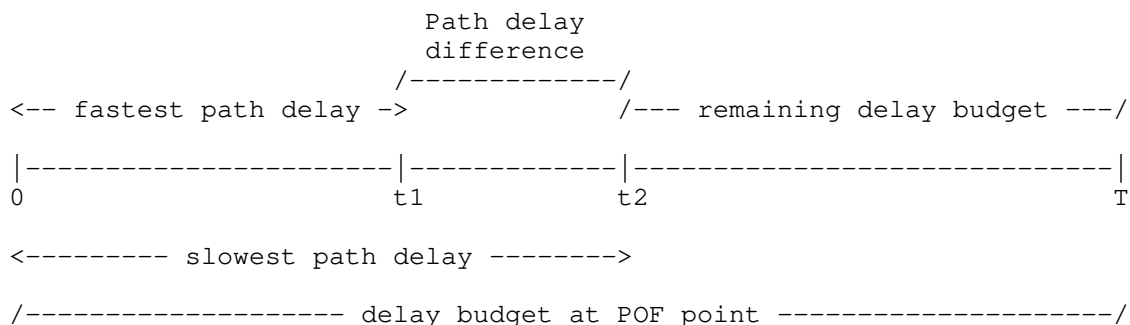


Figure 3: Delay Budget Relations at the POF Point

4.4. The Advanced POF Algorithm

In network scenario where the remaining delay budget of a flow at the POF point is smaller than "POFMaxDelay" time the basic method needs extensions.

The issue is that packets on the longest path cannot be buffered in order to keep delay budget of the flow. It must be noted that such a packet (i.e., forwarded over the longest path) needs no buffering as it is the "last chance" to deliver a packet with a given sequence number. This is because all replicas already must be arrived via shorter path(s).

The advanced POF algorithm needs two extensions of the basic POF algorithm:

- * to identify the received packet's path at the POF location and
- * to make the value of "POFMaxDelay" for buffered packets path dependent ("POFMaxDelay_i", where "i" notes the path the packet has used).

By identifying the path of a given frame, the POF algorithm can use this information to select what predefined time "POFMaxDelay_i" to apply for the buffered frame/packet. So, in the advanced POF algorithm "POFMaxDelay" is an array, that contains the predefined and path specific buffering time for each redundant path of a flow. Values in the "POFMaxDelay" array are engineered to fulfill the delay budget requirement.

The method for identification of the packet's path at the POF location depends on the network scenario. It can be implemented via various techniques, for example using ingress interface information, encoding the path in the packet itself (e.g., replication functions can set different FlowID per egress what can be used as a PathID), or in other means. Method for identification of the packet's path is out of scope in this document.

Note: in case of using the advanced POF algorithm it might be advantageous to combine PEF and POF locations in the DetNet network, as it can simplify the method used for identification of the packet's path at the POF location.

4.5. Further enhancements of POF algorithms

POF algorithms can be further enhanced by distinguishing the case of initialization from normal operation at the price of more states and more sophisticated implementation. Such enhancements could for example react better after some failure scenarios (e.g., complete outage of all paths of a DetNet flow) and may be dependent on the PEF implementation.

The challenge for POF initialization is that for example after a reset it is not known whether the first received packet is in-order or out-of-order. The original initialization (see before) considers the first packet as in-order, so out-of-order packet(s) during "POFMaxTime"/"POFMaxTime_path_i" time - after the first packet was received - may not be corrected. Motivation behind such an initialization is POF implementation simplicity.

A possible enhancement of POF initialization works as follows:

- * After a reset all received packets are buffered with their predefined timer ("POFMaxTime"/"POFMaxTime_path_i").
- * No basic/advanced POF rules are applied until the first timer expires.
- * When the first timer expires the packet with lowest seq_num in buffer is selected, forwarded, and "POFLastSent" is set with its seq_num.
- * The basic/advanced POF rules are applied for the packet(s) in the buffer and the subsequently received packets.

4.6. Selecting and using the POF algorithm

The selection of the POF algorithm depends on the network scenario and the remaining delay budget of a flow. Using POF and calculating its parameters require proper design. Knowing the path delay difference is essential for the POF algorithms described here. Failure scenarios breaking the design assumptions may impact the result of POF (e.g., packet received out of the expected worst-case delay window - calculated based on the path delay difference - may result in unwanted out-of-order delivery).

In DetNet scenarios there is always an Elimination function before the POF (therefore duplicates are not considered by the POF). Implementing them together in the same node allows POF to consider PEF events/states during the re-ordering. For example, under normal circumstances the difference of sequence number in consecutive

packets is bounded due to the history window of PEF. However, in some scenarios (e.g., reset of sequence number) the difference can be much larger than the history window size.

5. Control and Management Plane Parameters for POF

POF algorithms needs setting of the following parameters:

- * Basic POF

- "POFMaxDelay"
- "POFTakeAnyTime"

- * Advanced POF

- "POFMaxDelay_i"
- "POFTakeAnyTime"
- Network path identification related configuration(s)

Note, that in a proper design "POFTakeAnyTime" must be always larger than "POFMaxDelay".

6. Security Considerations

PREOF related security considerations (including POF) are described in section 3.3 of [RFC9055]. There are no additional POF related security considerations originating from this document.

7. IANA Considerations

This document makes no IANA requests.

8. Acknowledgements

Authors extend their appreciation to Gyorgy Miklos for his insightful comments and productive discussion that helped to improve the document.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.
- [RFC9055] Grossman, E., Ed., Mizrahi, T., and A. Hacker, "Deterministic Networking (DetNet) Security Considerations", RFC 9055, DOI 10.17487/RFC9055, June 2021, <<https://www.rfc-editor.org/info/rfc9055>>.

9.2. Informative References

- [IEEE8021CB] IEEE, "IEEE Standard for Local and metropolitan area networks -- Frame Replication and Elimination for Reliability", DOI 10.1109/IEEESTD.2017.8091139, October 2017, <https://standards.ieee.org/standard/802_1CB-2017.html>.
- [IEEEP8021CBcv] Kehrer, S., "FRER YANG Data Model and Management Information Base Module", IEEE P802.1CBcv /D1.2 P802.1CBcv, March 2021, <<https://www.ieee802.org/1/files/private/cv-drafts/d1/802-1CBcv-d1-2.pdf>>.

Authors' Addresses

Balázs Varga (editor)
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary
Email: balazs.a.varga@ericsson.com

János Farkas
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary
Email: janos.farkas@ericsson.com

Stephan Kehrer
Hirschmann Automation and Control GmbH
Stuttgarter Strasse 45-51.
72654 Neckartenzlingen
Germany
Email: Stephan.Kehrer@belden.com

Tobias Heer
Hirschmann Automation and Control GmbH
Stuttgarter Strasse 45-51.
72654 Neckartenzlingen
Germany
Email: Tobias.Heer@belden.com

DetNet
Internet-Draft
Intended status: Informational
Expires: 23 July 2022

B. Varga
J. Farkas
G. Mirsky
Ericsson
19 January 2022

Deterministic Networking (DetNet): OAM Functions for The Service Sub-
Layer
draft-varga-detnet-service-sub-layer-oam-02

Abstract

Operation, Administration, and Maintenance (OAM) tools are essential for a deterministic network. The DetNet architecture [RFC8655] has defined two sub-layers: (1) DetNet service sub-layer and (2) DetNet forwarding sub-layer. OAM mechanisms exist for the DetNet forwarding sub-layer. Nonetheless, OAM for the service sub-layer might require new extensions to the existing OAM protocols. This draft presents an analysis of OAM procedures for the DetNet service sub-layer functions.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 23 July 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights

and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
2.1. Terms Used in This Document	3
2.2. Abbreviations	3
2.3. Requirements Language	4
3. DetNet Service Sub-layer OAM Challenges	4
3.1. Illustrative example	4
3.2. DetNet Service Sub-layer Specifics for OAM	5
3.3. Information Needed during DetNet OAM Packet Processing .	6
3.4. A Possible Format of DetNet Associated Channel Header (d-ACH)	6
4. Requirements on OAM for DetNet Service Sub-layer	6
5. DetNet PING	6
5.1. Overview	7
5.2. OAM processing at the DetNet service sub-layer	7
5.2.1. Relay node with PRF	7
5.2.2. Relay node with PEF	8
5.2.3. Relay node with POF	9
5.2.4. Relay node without PREOF	9
6. Security Considerations	10
7. IANA Considerations	10
7.1. DetNet MPLS OAM Flags Registry	10
8. Acknowledgements	10
9. References	10
9.1. Normative References	10
9.2. Informative References	11
Authors' Addresses	12

1. Introduction

The DetNet Working Group has defined two sub-layers: (1) DetNet service sub-layer, at which a DetNet service (e.g., service protection) is provided and (2) DetNet forwarding sub-layer, which optionally provides resource allocation for DetNet flows over paths provided by the underlying network. In [RFC8655] new DetNet-specific functions have been defined for the DetNet service sub-layer, namely PREOF (a collective name for Packet Replication, Elimination, and Ordering Functions).

Framework of Operations, Administration and Maintenance (OAM) for Deterministic Networking (DetNet) is described in [I-D.ietf-detnet-oam-framework]. OAM for the DetNet MPLS data plane is described in [I-D.ietf-detnet-mpls-oam] and OAM for the DetNet IP data plane is described in [I-D.ietf-detnet-mpls-oam].

This draft has been submitted as an individual contribution to OAM discussions, in particular, to kick-off Working Group discussions on introducing OAM functions for the DetNet service sub-layer. It is also up to the Working Group discussions to which draft parts of this contribution may go, if any.

The OAM functions for the DetNet service sub-layer allow, for example, to recognize/discover DetNet relay nodes, to get information about their configuration, and to check their operation or status. Furthermore, the OAM functions for the DetNet service sub-layer need to meet new challenges (see section Section 3) and requirements (see section Section 4) introduced by PREOF.

An approach described in this draft introduces a new OAM shim layer to achieve OAM for the DetNet service sub-layer. In the rest of the draft, this approach is referred to as "DetNet PING", which is an in-band OAM approach, i.e., the OAM packets follow precisely the same path as the data packets of the corresponding DetNet flow(s) The OAM packets provide DetNet service sub-layer specific information, like:

- * Identity of a DetNet service sub-layer node.
- * Discover Ingress/Egress flow-specific configuration of a DetNet service sub-layer node.
- * Detect the status of the flow-specific service sub-layer function.

DetNet PING applies both to IP and MPLS DetNet data planes.

2. Terminology

2.1. Terms Used in This Document

This document uses the terminology established in the DetNet architecture [RFC8655], and the reader is assumed to be familiar with that document and its terminology.

2.2. Abbreviations

The following abbreviations are used in this document:

DetNet Deterministic Networking.

PEF Packet Elimination Function.

POF Packet Ordering Function.

PREOF Packet Replication, Elimination and Ordering Functions.

PRF Packet Replication Function.

2.3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. DetNet Service Sub-layer OAM Challenges

3.1. Illustrative example

This section introduces an example that is used to explain the DetNet Service Sub-layer OAM challenges. Figure 1 shows a DetNet flow on which PREOF functions are applied during forwarding from source to destination.

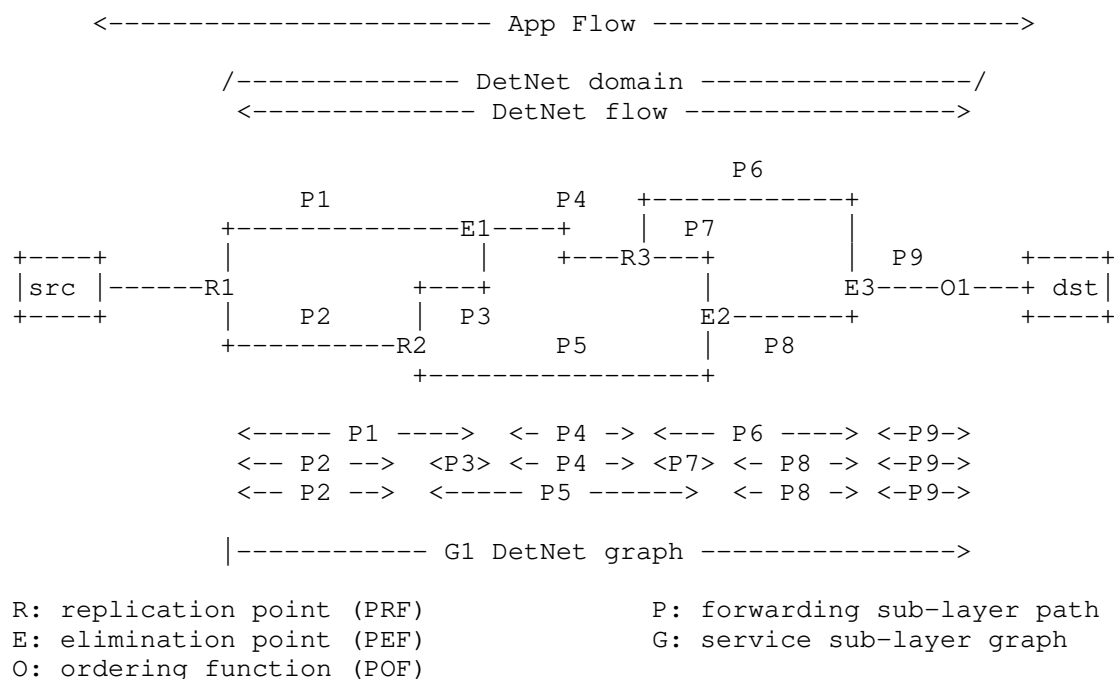


Figure 1: PREOF scenario in a DetNet network

DetNet service sub-layer nodes are interconnected by DetNet forwarding sub-layer paths. DetNet forwarding sub-layer path (e.g., P1 = R1->E1 path, P4 = E1->R3 path) may contain multiple transit nodes. A DetNet forwarding sub-layer path is used by a member flow and terminated by relay nodes (see [RFC8655] for relay node definition).

A DetNet service sub-layer graph includes all relay nodes and the interconnecting forwarding sub-layer paths. This graph can be also called as "PREOF graph" and it describes the compound flow as a whole.

3.2. DetNet Service Sub-layer Specifics for OAM

Several DetNet Service Sub-layer specifics have to be considered for OAM.

1. The service sub-layer graph is segmented into multiple parts, as forwarding sub-layer paths are terminated at DetNet relay nodes.

2. These are particular characteristics of DetNet PW:

1. PREOF acts as per-packet protection. PEF is a brand-new functionality at network layer, due to the per-packet merge action.
2. All paths are active and forward traffic. These paths may have a different number of hops.
3. Mandatory usage of a sequence number.

The above specifics have to be considered in combination with the requirement that DetNet OAM and DetNet data flows MUST receive the same treatment. OAM packets MUST follow precisely the same graph as the monitored DetNet flow(s).

3.3. Information Needed during DetNet OAM Packet Processing

This section collects some questions that have been already discussed by the DetNet WG and/or require further discussions by the WG. The section is structured in the form of a question list.

Question-1: Injecting OAM traffic in a DetNet flow? A DetNet data flow has a continuous Sequence Number. In order not to spoil that, the injected OAM packets require OAM-specific Sequence Number added. (See also Section 5.)

Question-2: How to process OAM packets by DetNet service sub-layer nodes? In order to cover the DetNet forwarding graph by OAM, PREOF has to be executed in an OAM specific manner (i.e., PREOF uses a separate SeqNum space for OAM. See details in Section 5.

Note: the question list is non-exhaustive.

3.4. A Possible Format of DetNet Associated Channel Header (d-ACH)

[Editor's note: The content of this section has been discussed and the outcome of the discussion has been documented in [I-D.ietf-detnet-mpls-oam].]

4. Requirements on OAM for DetNet Service Sub-layer

[Editor's note: The content of this section has been discussed and the outcome of the discussion has been documented in [I-D.ietf-detnet-oam-framework].]

5. DetNet PING

5.1. Overview

The "DetNet PING" approach uses two types of OAM packets: (1) DetNet-Echo-Request and (2) DetNet-Echo-Reply. Their encapsulation is identical to that of the corresponding DetNet data flow, so they follow precisely the same path as the packets of the corresponding DetNet data flow. They target DetNet service sub-layer entities, so they may not be recognized as OAM packets by entities not implementing DetNet service sub-layer for a packet flow (e.g., transit nodes). Other entities treat them as packets belonging to the corresponding DetNet data flow.

The following relay node roles can be distinguished:

1. DetNet PING originator node,
2. Intermediate DetNet service sub-layer node,
3. DetNet PING targeted node.

An originator node sends (generates) DetNet-Echo-Request packet(s). DetNet-Echo-Request packet contains an OAM specific "PINGSeqNum", which can be used by the DetNet service sub-layer of relay nodes. Note that "PINGSeqNum" is originator specific.

An intermediate DetNet service sub-layer node executes DetNet flow-specific service sub-layer functionality. Packet processing may be done in an OAM specific manner (see details in Section 5.2).

A targeted node answers with DetNet-Echo-Reply packet for each DetNet-Echo-Request. DetNet-Echo-Reply packet provides DetNet service sub-layer specific information on (i) identities of DetNet service sub-layer node (e.g., Node-ID, local Flow-ID), (ii) ingress/egress flow related configuration (e.g., in/out member flow specific information (including forwarding sub-layer specifics)), and (iii) status of service sub-layer function (e.g., local PxP-ID, Action-Type=x, operational status, value of key state variable(s), function related counters).

5.2. OAM processing at the DetNet service sub-layer

Detailed OAM packet processing rules of various DetNet relay nodes are described in the following sections.

5.2.1. Relay node with PRF

A DetNet relay node with PRF processes DetNet OAM packets in a stateless manner.

If the relay node with PRF is the target of a DetNet-Echo-Request packet, then the DetNet-Echo-Request packet MUST NOT be further forwarded, and a DetNet Echo-Reply packet MUST be generated. If the relay node with PRF is not the target of a DetNet Echo-Request packet, then the DetNet Echo-Request packet MUST be sent over all DetNet flow specific member flow paths (i.e., it is replicated).

A DetNet Echo-Reply packet MUST contain the following information:

- * Identities related to the DetNet service sub-layer node (e.g., Node-ID, local Flow-ID),
- * Ingress/Egress flow related configuration (e.g., in/out member flow specific information (including forwarding sub-layer specifics)),
- * Status of service sub-layer function (e.g., local PRF-ID, Action-Type=Replication, operational status, value of the flow related key state variable (e.g., "GenSeqNum" in [IEEE8021CB])).

A DetNet Echo-Reply packet MAY contain the following information:

- * PRF related local counters.

5.2.2. Relay node with PEF

A DetNet relay node with PEF processes DetNet OAM packets in a stateful manner.

If the relay node with PEF is the target of DetNet-Echo-Request packet, then the DetNet Echo-Request packet MUST NOT be further forwarded and an DetNet Echo-Reply packet MUST be generated. If the relay node with PEF is not the target of DetNet Echo-Request packet, then elimination MUST be executed on the DetNet Echo-Request packet(s) using the OAM specific "PINGSeqNum" in the packet. So only a single DetNet Echo-Request packet is forwarded and all further replicas (having the same originator's sequence number) MUST be discarded.

Note, PEF MAY use a simplified elimination algorithm for DetNet Echo-Request packets (e.g., "MatchRecoveryAlgorithm" in [IEEE8021CB]) as OAM is a slow protocol.

A DetNet-Echo-Reply packet MUST contain the following information:

- * Identities related to the DetNet service sub-layer node (e.g., Node-ID, local Flow-ID),

- * Ingress/Egress flow related configuration (e.g., in/out member flow specific information (including forwarding sub-layer specifics)) ,
- * Status of service sub-layer function (e.g., local PEF-ID, Action-Type=Elimination, operational status, value of the flow related key state variable (e.g., "RecovSeqNum" in [IEEE8021CB])).

A DetNet Echo-Reply packet MAY contain the following information:

- * PEF-related local counters.

5.2.3. Relay node with POF

A DetNet relay node with POF processes DetNet OAM packets in a stateless manner.

If the relay node with POF is the target of DetNet Echo-Request packet, then the DetNet Echo-Request packet MUST NOT be further forwarded and a DetNet Echo-Reply packet MUST be generated. If the relay node with POF is not the target of DetNet-Echo-Request packet, then the DetNet Echo-Request packet(s) MUST be forwarded without any POF-specific action.

A DetNet Echo-Reply packet MUST contain the following information:

- * Identities of the DetNet service sub-layer node (e.g., Node-ID, local Flow-ID),
- * Ingress/Egress flow related configuration (e.g., in/out member flow specific information (including forwarding sub-layer specifics)) ,
- * Status of service sub-layer function (e.g., local POF-ID, Action-Type=Ordering, operational status, value of the flow related key state variable (e.g., "POFLastSent" in [I-D.varga-detnet-pof])).

A DetNet Echo-Reply packet MAY contain the following information:

- * POF-related local counters.

5.2.4. Relay node without PREOF

A DetNet relay node without PREOF processes DetNet OAM packets in a stateless manner.

If the relay node without PREOF is the target of DetNet Echo-Request packet, then the DetNet Echo-Request packet MUST NOT be further forwarded and an DetNet Echo-Reply packet MUST be generated. If the relay node without PREOF is not the target of DetNet-Echo-Request packet, then the DetNet-Echo-Request packet(s) MUST be forwarded (as any data packets of the related DetNet flow).

A DetNet Echo-Reply packet MUST contain the following information:

- * Identities of the DetNet service sub-layer node (e.g., Node-ID, local Flow-ID),
- * Ingress/Egress flow-related configuration (e.g., in/out member flow specific information (including forwarding sub-layer specifics)) .

6. Security Considerations

Tbd.

7. IANA Considerations

7.1. DetNet MPLS OAM Flags Registry

[Editor's note: The content of this section has been discussed and the outcome of the discussion has been documented in [I-D.ietf-detnet-mpls-oam].]

8. Acknowledgements

Authors extend their appreciation to Janos Szabo and Gyorgy Miklos for their insightful comments and productive discussion that helped to improve the document.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, DOI 10.17487/RFC4928, June 2007, <<https://www.rfc-editor.org/info/rfc4928>>.

- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.
- [RFC8964] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., Bryant, S., and J. Korhonen, "Deterministic Networking (DetNet) Data Plane: MPLS", RFC 8964, DOI 10.17487/RFC8964, January 2021, <<https://www.rfc-editor.org/info/rfc8964>>.

9.2. Informative References

- [I-D.ietf-detnet-ip-oam]
Mirsky, G., Chen, M., and D. Black, "Operations, Administration and Maintenance (OAM) for Deterministic Networks (DetNet) with IP Data Plane", Work in Progress, Internet-Draft, draft-ietf-detnet-ip-oam-03, 19 September 2021, <<https://www.ietf.org/archive/id/draft-ietf-detnet-ip-oam-03.txt>>.
- [I-D.ietf-detnet-mpls-oam]
Mirsky, G., Chen, M., Varga, B., and J. Farkas, "Operations, Administration and Maintenance (OAM) for Deterministic Networks (DetNet) with MPLS Data Plane", Work in Progress, Internet-Draft, draft-ietf-detnet-mpls-oam-06, 10 December 2021, <<https://www.ietf.org/archive/id/draft-ietf-detnet-mpls-oam-06.txt>>.
- [I-D.ietf-detnet-oam-framework]
Mirsky, G., Theoleyre, F., Papadopoulos, G. Z., Bernardos, C. J., Varga, B., and J. Farkas, "Framework of Operations, Administration and Maintenance (OAM) for Deterministic Networking (DetNet)", Work in Progress, Internet-Draft, draft-ietf-detnet-oam-framework-05, 14 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-detnet-oam-framework-05.txt>>.

[I-D.varga-detnet-pof]

Varga, B., Farkas, J., Kehrler, S., and T. Heer,
"Deterministic Networking (DetNet): Packet Ordering
Function", Work in Progress, Internet-Draft, draft-varga-
detnet-pof-02, 22 October 2021,
<<https://www.ietf.org/archive/id/draft-varga-detnet-pof-02.txt>>.

[IEEE8021CB]

IEEE, "IEEE Standard for Local and metropolitan area
networks -- Frame Replication and Elimination for
Reliability", DOI 10.1109/IEEESTD.2017.8091139, October
2017,
<https://standards.ieee.org/standard/802_1CB-2017.html>.

Authors' Addresses

Balázs Varga
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary

Email: balazs.a.varga@ericsson.com

János Farkas
Ericsson
Budapest
Magyar Tudosok krt. 11.
1117
Hungary

Email: janos.farkas@ericsson.com

Greg Mirsky
Ericsson

Email: gregimirsky@gmail.com

DETNET
Internet-Draft
Intended status: Standards Track
Expires: 27 April 2022

Q. Xiong
ZTE Corporation
October 2021

The Requirements for Wide-area IP Deterministic Networking
draft-xiong-detnet-wide-area-ip-requirements-00

Abstract

In wide-area IP networks, more requirements need to be taken into considerations such as differentiated DetNet QoS of multiple services, high-efficiency of resources utilization and routes steering, integration of large-scale heterogeneous network and guarantees of multiple dynamic deterministic flows. This document describes the requirements in wide-area applications and proposes the solution with deterministic resources, routes and QoS.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 4 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions used in this document	4
2.1. Terminology	4
2.2. Requirements Language	4
3. Requirements for Wide-area IP Deterministic Networking	4
3.1. Differentiated DetNet QoS of Multiple Services	4
3.2. Integration of Large-scale Heterogeneous Network	5
3.3. Efficiency of Resources Utilization and Routes Steering	6
3.4. Guarantees of Multiple Dynamic Deterministic Flows	6
4. Solutions Considerations of Wide-area IP Deterministic Networking	7
4.1. The Deterministic Resources	7
4.2. The Deterministic Routes	7
4.3. The Deterministic QoS	8
5. Security Considerations	8
6. Acknowledgements	8
7. IANA Considerations	8
8. Normative References	8
Author's Address	9

1. Introduction

5G network is oriented to the internet of everything. In addition to the Enhanced Mobile Broadband (eMBB) and Massive Machine Type Communications (mMTC) services, it also supports the Ultra-reliable Low Latency Communications (uRLLC) services. The uRLLC services cover the industries such as intelligent electrical network, intelligent factory, internet of vehicles, industry automation and other industrial internet scenarios, which is the key demand of digital transformation of vertical domains. These uRLLC services demand SLA guarantees such as low latency and high reliability and other deterministic and precise properties.

For the intelligent electrical network, there are deterministic requirements for communication delay, jitter and packet loss rate. For example, in the electrical current difference model, a delay of

3~10ms and a jitter variation is no more than 100us are required. The isolation requirement is also important. For example, the automatic operation, control of a process, isochronous data and low priority service need to meet the requirements of hard isolation. In addition to the requirements of delay and jitter, the differential protection (DP) service needs to be isolated from other services.

The industrial internet is the key infrastructure that coordinate various units of work over various system components, e.g. people, machines and things in the industrial environment including big data, cloud computing, Internet of Things (IOT), Augment Reality (AR), industrial robots, Artificial Intelligence (AI) and other basic technologies. For example, automation control is one of the basic application and the the core is closed-loop control system. The control process cycle is as low as millisecond level, so the system communication delay needs to reach millisecond level or even lower to ensure the realization of precise control. There are three levels of real-time requirements for industrial interconnection: factory level is about 1s, and process level is 10~100ms, and the highest real-time requirement is motion control, which requires less than 1ms.

The applications in 5G networks demand much more deterministic and precise properties. But traditional Ethernet, IP and MPLS networks which is based on statistical multiplexing provides best-effort packet service and offers no delivery and SLA guarantee. The deterministic forwarding can only apply to flows with such well-defined characteristics as periodicity and burstiness. Technologies to provide deterministic service has been proposed to provide bounded latency and jitter based on a best-effort packet network. IEEE 802.1 Time-Sensitive Networking (TSN) has been proposed to provide bounded latency and jitter in L2 LAN networks. According to [RFC8655], Deterministic Networking (DetNet) operates at the IP layer and delivers service which provides extremely low data loss rates and bounded latency within a network domain.

The deterministic networks not only need to offer the Service Level Agreements (SLA) guarantees such as low latency and jitter, low packet loss and high reliability, but also need to support the precise services such as flexible resource allocation and service isolation. However, under the existing IP network architecture with statistical multiplexing characteristics, the existing deterministic technologies are facing large scale number of nodes and long-distance transmission, traffic scheduling, dynamic flows, and other controversial issues especially in Wide Area Network (WAN) applications.

In wide-area IP networks, more requirements need to be taken into considerations such as differentiated DetNet QoS of multiple services, high-efficiency of resources utilization and routes steering, integration of large-scale heterogeneous network and guarantees of multiple dynamic deterministic flows. This document describes the requirements in wide-area applications and proposes the solution with deterministic resources, routes and QoS.

2. Conventions used in this document

2.1. Terminology

The terminology is defined as [RFC8655].

2.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Requirements for Wide-area IP Deterministic Networking

3.1. Differentiated DetNet QoS of Multiple Services

As defined in [RFC8655], the DetNet QoS can be expressed in terms of : Minimum and maximum end-to-end latency, bounded jitter (packet delay variation), packet loss ratio and an upper bound on out-of-order packet delivery. As described in [RFC8578], DetNet applications differ in their network topologies and specific desired behavior and different services requires differentiated DetNet QoS. In the WAN scenarios, multiple services with differentiated DetNet QoS is co-existed in the same DetNet network. The classification of the deterministic flows is should be taken into considerations. It is required to provide Latency, bounded jitter and packet loss dynamically and flexibly in all scenarios for each characterizd flow.

As the Figure 1 shown, the services is divided into 4 levels and level 1~3 is the DetNet flows and level-4 is non-DetNet flow. DetNet applications and DetNet QoS is differentiated within each level.

Item	Level-1	Level-2	Level-3	Level-4
Applications Examples	Industrial	VR/AR	Audio and Video	Broadcast
DetNet QoS	Ultra-low latency and jitter	Low latency and jitter	Low latency	Best Effort

Figure 1: Figure 1: The classification of multiple services

Moreover, different DetNet services is required to tolerate different percentage of packet loss ratio such as 99.9%, 99.99%, 99.999%, and so on. It is also required to provide service isolation. In some scenarios, such as intelligent electrical network, the isolation requirements are very important. For example, the automatic operation or control of a process or isochronous data and service with different priorities need to meet the requirements of hard isolation. In addition to the requirements of delay and jitter, the differential protection (DP) service needs to be isolated from other services and hard isolated tunnel is required.

3.2. Integration of Large-scale Heterogeneous Network

In WAN application, large-scale number of nodes and long-distance transmission in the network will lead to latency and jitter, such as increasing transmission latency, jitter and packet loss. It is to required reduce the scale of the network topology by establishing cutthrough channels. The existing technologies such as FlexE and SR tunnels should be taken into consideration. And multiple capabilities is also provided by the nodes and links within the network topology such as FlexE tunnels, TSN sub-network and IP/MPLS/SRv6 tunnels. It is required to integrate the multi-capability resources to achieve the optimal DetNet QoS.

Another option is to divide the network into several domains and segments. And the deadline of latency and jitter of each domain and segment should be determined and controlled. It is required to control the DetNet QoS at the inter-domain boundary nodes and achieve the end-to-end latency, bounded jitter and packet loss ratio across.

3.3. Efficiency of Resources Utilization and Routes Steering

Traditional Ethernet, IP and MPLS networks which is based on statistical multiplexing provides best-effort packet service and offers no delivery and SLA guarantee. As described in [RFC8655], the primary technique by which DetNet achieves its QoS is to allocate sufficient resources. But it can not be achieved by not sufficient resource which can be allocated due to practical and cost reason. So it is required to achieve the high-efficiency of resources utilization when provide the DetNet service.

Network resources include nodes, links, ports, bandwidth, queues, etc. The congestion control, shaping and queue scheduling and other traffic mechanisms which have been proposed in IEEE 802.1 TSN such as IEEE802.1Qbv, IEEE802.1Qch, IEEE802.1Qav, IEEE802.1Qcr and so on. Heterogeneous resource should be used and unified and simplified resources mechanism under the selection of existing multiple technical methods to realize the elastic of deterministic capability.

Resource classification and modeling is required along with the explicit path with more SLA guarantee parameters like bandwidth, latency, jitter, packet loss and so on. On the basic of the resources, the steering path and routes for deterministic flows should be programmed before the flows coming and able to provide SLA capability. And the routes should be considered to be established in distributed and centralized control Plane.

3.4. Guarantees of Multiple Dynamic Deterministic Flows

As described in [RFC8557], deterministic forwarding can only apply to flows with such well-defined characteristics as periodicity and burstiness. As defined in DetNet architecture [RFC8655], the traffic characteristics of an App-flow can be CBR (constant bit rate) or VBR (variable bit rate) of L1, L2 and L3 layers (VBR takes the maximum value when reserving resources). But the current scenarios and technical solutions only consider CBR flow, without considering the coexistence of VBR and CBR, the burst and aperiodicity of flows. The operations such as shaping or scheduling have not been specified. Even TSN mechanisms are based on a constant and forecastable traffic characteristics.

It will be more complicated in WAN applications where much more flows coexist and the traffic characteristics is more dynamic. It is required to offer reliable delivery and SLA guarantee for dynamic flows. For example, periodic flow and aperiodic flow (including micro burst flow, etc.), CBR and VBR flow, flow with different periods or phases, etc. When the network needs to forward these deterministic flows at the same time, it must solve the problems of

time window selection, queue processing and aggregation of multiple flows. It is required to classify the dynamic deterministic flows and map them into different virtual topologies to limit the number of the concurrent flows and reduce the micro bursts.

4. Solutions Considerations of Wide-area IP Deterministic Networking

4.1. The Deterministic Resources

As defined in RFC8655, the resource allocation is one of the techniques to achieve the DetNet QoS. Network resources include nodes, links, ports, bandwidth, queues, etc. The deterministic resources require planning and arrangement of network resources, resources modeling, resource allocation and reservation, resource isolation and resource scheduling, etc. In order to meet the requirements of deterministic service, resources need to be classified, including ultra-low delay resources, low delay resources, low jitter resources, etc.

Deterministic resources guarantee the delay, jitter and other requirements of deterministic services by reserving resources for flows. If the network resources are sufficient, congestion and packet loss can be eliminated to meet the requirements of low delay jitter. If the network resources are insufficient, congestion control, queue mechanisms of deterministic flows need to be carried out. The nodes with different queue mechanisms provide different latency and bounded jitter. Moreover, network resources could be reconstructed to provide ultra-low latency, for example, L1 layer resources could be used to provide cutthrough channels, FlexE pipes, etc.

4.2. The Deterministic Routes

The deterministic routes is based on the provision of deterministic resources. The deterministic routes refers to the requirements to select the network routes for the deterministic flows to guarantee the stability of the routing at least during the packets transmission, and the path will not change within the real-time change of network topology. Moreover, the deterministic routes should provide the capability including the latency, jitter and packet loss ratio.

Routes generally perform forwarding function including receiving the incoming packets and forwarding the packets to a Router based on the header information and a forwarding information base. It is necessary to provide pre-routes with SLA capability and generate endogenous deterministic routing with deterministic capability. The deterministic routes perform the functions of forwarding and QoS

guarantee at the same time. The types of deterministic routes can be classified into ultra-low delay routes, low delay routes, low jitter routes, and so on. There can also has replication routes and aggregation routes.

The mechanisms of path establishment include traffic engineering technology (MPLS-TE, SR-TE, static configuration, etc.), IGP technology, etc. Explicit strict routing can guarantee the delay jitter and other requirements of services. Loose routing only generates some endogenous deterministic routes, and other routes still need forwarding and scheduling, such as dynamic resource-aware routing and queue scheduling.

4.3. The Deterministic QoS

The deterministic QoS is to arrange and schedule the deterministic flows on the basis of providing deterministic resources and routes, so as to control of each flow and meet the DetNet QoS goals.

The scheduling and control include the classification of the deterministic flows, queue scheduling mechanism for each class of deterministic flow, deterministic shaping at boundary nodes, limiting the number of concurrent flows and reducing micro bursts, mapping the dynamic concurrent flows into different virtual topologies. Moreover, flow aggregation is performed at the aggregation node to reduce flow state maintenance and replication or elimination is performed at the relay node to achieve reliability.

If the deterministic flows crosses multiple domains, the end-to-end latency is the sum of delay from all domains. It is required to control the deadline delay of each domain. Moreover, bounded jitter (packet delay variation) should be adjusted and scheduled at the inter-domain boundary nodes.

5. Security Considerations

TBA

6. Acknowledgements

TBA

7. IANA Considerations

TBA

8. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8557] Finn, N. and P. Thubert, "Deterministic Networking Problem Statement", RFC 8557, DOI 10.17487/RFC8557, May 2019, <<https://www.rfc-editor.org/info/rfc8557>>.
- [RFC8578] Grossman, E., Ed., "Deterministic Networking Use Cases", RFC 8578, DOI 10.17487/RFC8578, May 2019, <<https://www.rfc-editor.org/info/rfc8578>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.

Author's Address

Quan Xiong
ZTE Corporation
No.6 Huashi Park Rd
Wuhan
Hubei, 430223
China

Email: xiong.quan@zte.com.cn

IDR
Internet-Draft
Intended status: Standards Track
Expires: November 28, 2021

Q. Xiong
H. Wu
ZTE Corporation
May 27, 2021

BGP Flow Specification for DetNet Flow Mapping
draft-xiong-idr-detnet-flow-mapping-00

Abstract

This document proposes extensions to BGP Flow Specification for the flow mapping of Deterministic Networking (DetNet) when interconnected with IEEE 802.1 Time-Sensitive Networking (TSN). The BGP flowspec is used for the filtering of the packets that match the DetNet networks and the mapping between TSN streams and DetNet flows in the control plane.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on November 28, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
2.1. Terminology	3
2.2. Requirements Language	3
3. The Flow Mapping of DetNet	3
4. BGP Extensions for Flow Specification Encoding	4
4.1. Filtering Rules for TSN Streams	4
4.2. Traffic Action for TSN Streams	5
4.3. Filtering Rules for DetNet Flows	6
4.4. Traffic Action for DetNet Flows	7
5. Security Considerations	8
6. Acknowledgements	8
7. IANA Considerations	8
8. Normative References	8
Authors' Addresses	9

1. Introduction

[RFC8655] specifies the architecture of Deterministic Networking (DetNet), which provide a capability for the delivery of data flows with extremely low packet loss rates and bounded end-to-end delivery latency. DetNet-enabled end systems and DetNet nodes can be interconnected by sub-networks, i.e., Layer 2 technologies such as IEEE 802.1 Time-Sensitive Networking (TSN).

As defined in [RFC8655], the DetNet IP and MPLS flows can be carried over TSN sub-networks. DetNet needs to be mapped to the sub-networks technology used to interconnect DetNet nodes. For example, a TSN node may be used to interconnect DetNet-aware nodes, and these DetNet nodes can map DetNet flows to TSN streams. When the Detnet provide the deterministic service for the TSN end system, a DetNet edge node may be used to interconnect the TSN end system, and the DetNet nodes can map the TSN streams to DetNet flows.

As described in [RFC8938], one of the primary requirements of the DetNet Controller Plane is restricting flows to IEEE 802.1 TSN and the requirement could use the centralized network management provisioning mechanisms such as BGP protocol. As defined in [RFC8955], the Flow Specifications for BGP is an n-tuple consisting of several matching criteria which is comprised of traffic filtering rules and is associated with actions that can be applied to the traffic flows. The DetNet edge nodes can provide the capability to process the traffic including classifying, shaping, rate limiting,

filtering, and redirecting packets based on the policies configured by the BGP Flow Specification.

This document proposes extensions to BGP Flow Specification for the interconnection of DetNet and TSN. The BGP flowspec is used for the filtering of the packets that match the DetNet networks and the mapping between TSN streams and DetNet flows in the control plane.

2. Conventions used in this document

2.1. Terminology

The terminology is defined as [RFC8655], [RFC8938], and [RFC8955].

2.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. The Flow Mapping of DetNet

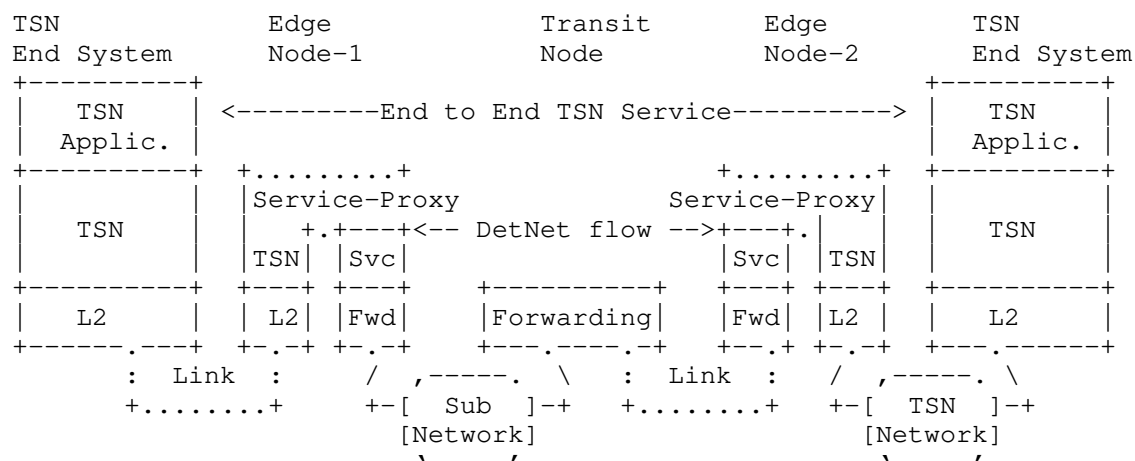
As described in [I-D.ietf-detnet-tsn-vpn-over-mpls], TSN networks can be interconnected over a DetNet MPLS Network. And as discussed in [I-D.ietf-detnet-ip-over-tsn] and [I-D.ietf-detnet-mpls-over-tsn], DetNet IP or MPLS networks can be operating over a TSN sub-network. The mapping between TSN Streams and DetNet flows is required for the service proxy function at DetNet Edge nodes. And the mapping table can be configured and maintained in the control plane. When a DetNet Edge Node receives a packet, it MUST identify and check whether such flow is present in its mapping table and decide to drop (when not match) or to forward the packet (when match) to the associated service. 1:1 and N:1 mapping (aggregating multiple TSN Streams in a single DetNet flow) MUST be supported.

As Figure 1 shows, it is required to configure the identification information when mapping received TSN Streams to the DetNet flows at Edge Node-1. Mechanisms and Parameters of TSN stream identification (e.g., Mask-and-Match Stream identification) defined in [IEEE8021CB] and [IEEE8021CBdb] can be used for service proxy function. After the identification of the TSN stream, it need to map the packet to the DetNet flow information such as S-Label, d-CW when in DetNet MPLS data plane and handle the packet as defined in [RFC8964].

When the DetNet Edge Node-2 receives a DetNet flow, it MUST identify the DetNet flow-ID information such as IP 6-tuple in DetNet IP data

plane or S-Label and d-CW information in DetNet MPLS data plane. Then the Service proxy function need to map the DetNet flow-ID and flow related parameters to the associated TSN Stream IDs and streams related parameters.

As defined in [RFC8955], the nodes that applied a Flow Specification can filter the received packets according to the matching criteria and can forward the flows based on the associated actions. This document proposes extensions to BGP Flow Specification for the mapping of DetNet flows and TSN streams by using the traffic filtering rules to identify the packet and using the associated action to map the packet to the related service.



Flow Mapping:

|TSN N:1 DetNet|<----- DetNet ----->|DetNet 1:N TSN|

Figure 1: Flow Mapping in TSN over DetNet Network

4. BGP Extensions for Flow Specification Encoding

4.1. Filtering Rules for TSN Streams

As IEEE Std 802.1Q defined, a Stream ID is a 64-bit field that uniquely identifies a stream and can be generated by the system offering the stream, or possibly a device controlling that system. But it is not carried in the header of the TSN Stream. As defined in [IEEE8021CB] and [IEEE8021CBdb], five specific Stream identification functions are described: Null Stream identification, Source MAC and VLAN Stream identification, Active Destination MAC and VLAN Stream

identification, and IP Stream identification, and Mask-and-match Stream identification. It needs to examine the header of the streams such as destination_address, vlan_identifier, IP source address, IP destination address, DSCP, IP next protocol, source port, destination port and mac_service_data_unit.

As defined in [I-D.ietf-idr-flowspec-l2vpn], the Ethernet Layer 2 (L2) related fields have been covered by the L2 traffic filtering rules except the mac_service_data_unit in Mask-and-Match Stream identification. A mac_service_data_unit mask is defined to identify communication flows supported by various higher-layer protocols. This document proposes a new type in L2 components flowspec Type for TSN Streams.

Type TBD1 - Mac Service Data Unit

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match 6-octet Mac Service Data Unit field. Values are encoded as 6-octet quantities. op is encoded as specified in Section 4.2.1.1 of [RFC8955].

4.2. Traffic Action for TSN Streams

The action for an TSN traffic filtering flowspec is to accept the TSN streams that matches that particular rule and map the streams to the DetNet flows. The action for L3 traffic with extended communities types per [RFC8955] and [RFC8956] such as traffic-rate, traffic-marking, traffic-action, and redirect can be used for TSN to DetNet IP flow mapping.

The DetNet flow is identified by a S-Label and the DetNet Header consists of d-CW and F-Labels. The MPLS label related action for an TSN stream mapping to a DetNet MPLS network can use the Label-action defined in [I-D.ietf-idr-bgp-flowspec-label]. And the action for the sequence in d-CW field, this document specifies the following BGP extended community for TSN Streams as following shown.

type	extended community	encoding
TBD2	Sequence-action	bitmask

Table 1

The The Sequence-action extended community is shown as the Figure 2.

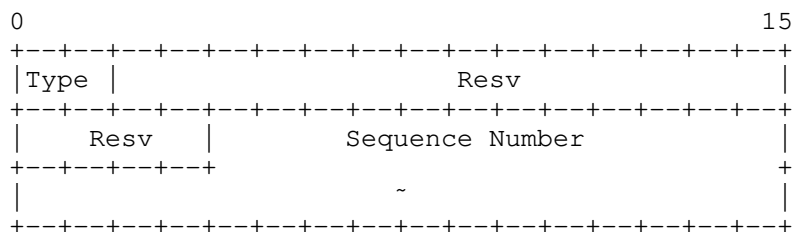


Figure 2: Sequence-action

Type: 2 bits, indicates the length of the sequence number:

0: 0 bits

1: 16 bits

2: 28 bits

Resv: 18 bits, reserved for future use. MUST be sent as zero and ignored on receipt.

Sequence Number: 28 bits, an unsigned value implementing the DetNet sequence number.

4.3. Filtering Rules for DetNet Flows

The L3 traffic filtering rules defined in [RFC8955] and [RFC8956] can be used for DetNet IP flow.

As defined in RFC8964, the MPLS-based DetNet data plane encapsulation consists of d-CW, S-Label and F-Labels. The MPLS label filtering rules have been defined in [I-D.ietf-idr-flowspec-mpls-match].

This document proposes a new community type in L3 components flowspec Type for DetNet MPLS flows.

Type TBD3 - d-CW

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match Sequence. Values are encoded as 4-octet quantities, where the four most significant bits are set to zero and ignored for matching and the 28 least significant bits contain the sequence value. op is encoded as specified in Section 4.2.1.1 of [RFC8955].

4.4. Traffic Action for DetNet Flows

The extended action for an DetNet traffic filtering flowspec is to accept the DetNet flows that matches that particular rule and map the flows to the TSN streams. This document specifies the following BGP extended communitiy as the following shown.

type	extended community	encoding
TBD4	TSN-action	bitmask

Table 2

The The TSN-action extended community is shown as the Figure 3.

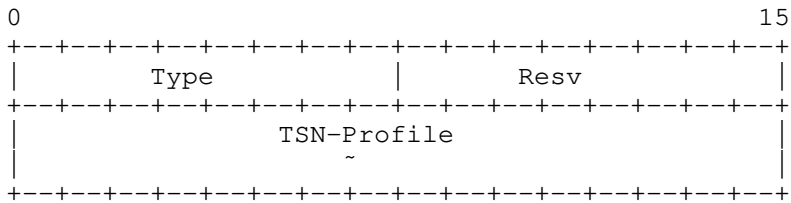


Figure 3: TSN-action

Type: 1-octet, indicates the type of TSN profiles. The value of the types is TBD:

Resv: 1-octet, reserved for future use. MUST be sent as zero and ignored on receipt.

TSN-profile: 4-octet, can be converted to the TSN Stream ID and stream related parameters and requirements as the following shown.

stream_handle: identifying the Stream to which the packet belongs in TSN networks.

sequence_number: identifying the order in which the packet was transmitted relative to other packets in the same Compound Stream in TSN networks.

traffic_scheduling: identifying the traffic scheduling mechanisms including traffic policy, queuing and forwarding methods in TSN networks.

5. Security Considerations

TBA

6. Acknowledgements

TBA

7. IANA Considerations

TBA

8. Normative References

[I-D.ietf-detnet-ip-over-tsn]

Varga, B., Farkas, J., Malis, A. G., and S. Bryant,
"DetNet Data Plane: IP over IEEE 802.1 Time Sensitive
Networking (TSN)", draft-ietf-detnet-ip-over-tsn-07 (work
in progress), February 2021.

[I-D.ietf-detnet-mpls-over-tsn]

Varga, B., Farkas, J., Malis, A. G., and S. Bryant,
"DetNet Data Plane: MPLS over IEEE 802.1 Time-Sensitive
Networking (TSN)", draft-ietf-detnet-mpls-over-tsn-07
(work in progress), February 2021.

[I-D.ietf-detnet-tsn-vpn-over-mpls]

Varga, B., Farkas, J., Malis, A. G., Bryant, S., and D.
Fedyk, "DetNet Data Plane: IEEE 802.1 Time Sensitive
Networking over MPLS", draft-ietf-detnet-tsn-vpn-over-
mpls-07 (work in progress), February 2021.

[I-D.ietf-idr-bgp-flowspec-label]

Liang, Q., Hares, S., You, J., Raszuk, R., and D. Ma,
"Carrying Label Information for BGP FlowSpec", draft-ietf-
idr-bgp-flowspec-label-01 (work in progress), December
2016.

[I-D.ietf-idr-flowspec-l2vpn]

Hao, W., Eastlake, D. E., Litkowski, S., and S. Zhuang,
"BGP Dissemination of L2 Flow Specification Rules", draft-
ietf-idr-flowspec-l2vpn-16 (work in progress), November
2020.

- [I-D.ietf-idr-flowspec-mpls-match]
Yong, L., Hares, S., Liang, Q., and J. You, "BGP Flow Specification Filter for MPLS Label", draft-ietf-idr-flowspec-mpls-match-01 (work in progress), December 2016.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <<https://www.rfc-editor.org/info/rfc8655>>.
- [RFC8938] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., and S. Bryant, "Deterministic Networking (DetNet) Data Plane Framework", RFC 8938, DOI 10.17487/RFC8938, November 2020, <<https://www.rfc-editor.org/info/rfc8938>>.
- [RFC8955] Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", RFC 8955, DOI 10.17487/RFC8955, December 2020, <<https://www.rfc-editor.org/info/rfc8955>>.
- [RFC8956] Loibl, C., Ed., Raszuk, R., Ed., and S. Hares, Ed., "Dissemination of Flow Specification Rules for IPv6", RFC 8956, DOI 10.17487/RFC8956, December 2020, <<https://www.rfc-editor.org/info/rfc8956>>.
- [RFC8964] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., Bryant, S., and J. Korhonen, "Deterministic Networking (DetNet) Data Plane: MPLS", RFC 8964, DOI 10.17487/RFC8964, January 2021, <<https://www.rfc-editor.org/info/rfc8964>>.

Authors' Addresses

Quan Xiong
ZTE Corporation
No.6 Huashi Park Rd
Wuhan, Hubei 430223
China

Email: xiong.quan@zte.com.cn

Haisheng Wu
ZTE Corporation
Nanjing, Jiangsu
China

Email: wu.haisheng@zte.com.cn