Workgroup: DETNET
Internet-Draft:
draft-eckert-detnet-mpls-tc-tcqf-02
Published: 19 May 2022
Intended Status: Standards Track
Expires: 20 November 2022
Authors: T. Eckert
Futurewei Technologies USA
S. Bryant
University of Surrey ICS
A. G. Malis
Malis Consulting

Deterministic Networking (DetNet) Data Plane - MPLS TC Tagging for Cyclic Queuing and Forwarding (MPLS-TC TCQF)

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 20 November 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 (informative)
2. Using TCQF in the DetNet Archticture and MPLS forwarding plane (informative)
3. MPLS T-CQF forwarding (normative)
  3.1. Configuration Data model and tag processing for MPLS TC TCQF
  3.2. Packet processing
  3.3. TCQF with label stack operations
  3.4. Ingres operations
4. TCQF Pseudocode (normative)
5. Operational considerations (informative)
  5.1. Controller plane computation of cycle mappings
6. Security Considerations
7. IANA Considerations
8. Changelog
9. References
  9.1. Normative References
  9.2. Informative References
Authors' Addresses

1. Introduction (informative)

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

CQF 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, CQF 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 Archticture 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.

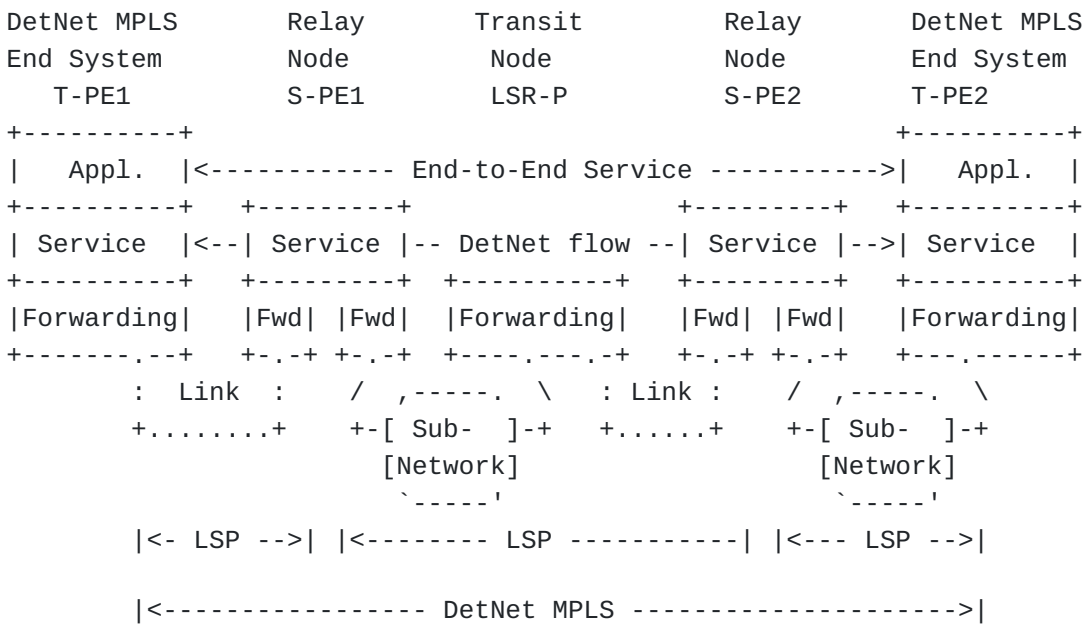


Figure 1: A DetNet MPLS Network

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 TCQF, a sequence of LSR and DetNet service node implements TC TCQF, 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 TCQF 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 CQF.

All LSR/Service node after the ingress node only have to map a received TCQF 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 TCQF 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 TCQF domain from T-PE1 to S-PE2, start a new TCQF 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 TCQF 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 TCQF (see Section 5.1).

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

Likewise, S-PE2 could perform the Packet Elimination Function without being a TCQF edge node as this most likely does not introduce any non-TCQF 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 TCQF egress/ingress, performing that shaping itself as an ingress function.

3. MPLS T-CQF forwarding (normative)

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

```
tcqf
+-- 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]

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

Figure 2: TCQF Configuration Data Model

3.2. Packet processing

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

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

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

tcqf.cycles is the number of cycles used across all interfaces in the TCQF 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 CQF domain.

The unit of tcqf.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 tcqf.cycle_clock_offset in units of nsec as follows. Let clock1 be a timestamp of the local reference clock for TCQF, at which cycle 1 starts, then:

tcqf.cycle_clock_offset = (clock1 mod (tcqf.cycle_time * tcqf.cycles))

The local reference clock of the LSR/router is expected to be synchronized with the neighboring LSR/router in TCQF domain. tcqf.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.

tcqf.if_config[oif] is optional per-interface configuration of TCQF parameters. tcqf.if_config[oif].cycle_clock_offset may be different from tcqf.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 tcqf.if_config[oif].cycle_clock_offset is used to indicate that the domain wide tcqf.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 tcqf.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.

tcqf_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 tcqf_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, TCQF 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. Ingres operations

The ingress LSR of a TCQF 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. TCQF Pseudocode (normative)

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

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

void inject_tcqf_pak(pak, cycle) {
  pak.context.iif = INTERNAL
  pak.context.tcqf_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.tcqf_cycle && // non TCQF packets cycle is 0
     tcqf.if_config[oif]) {    // TCQF enabled
    // Map tcqf_cycle iif to oif
    cycle = pak.context.tcqf_cycle
          = map_cycle(cycle,
            tcqf.if_config[oif].cycle_map[[iif])

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

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

// Started when TCQF is enabled on an interface
// dequeues packets from oif.cycleq
void send_tcqf(oif) {
  cycle = 1
  cc =  tcqf.cycle_time *
        tcqf.cycle_time
  o =   tcqf.cycle_clock_offset
```

```
  nextcyclestart = floor(tnow / cc) * cc + cc + o

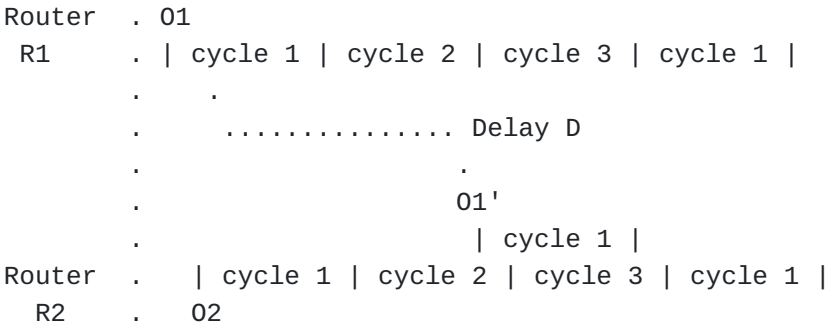
  while(1) {
    while(tnow < nextcyclestart) { }
    while(pak = dequeue(oif.cycleq(cycle)) {
      send(pak)
    }
    cycle = (cycle + 1) mod tcqf.cycles + 1
    nextcyclestart += tcqf.cycle_time
  }
}
```

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.




```
CT  = cycle_time
C   = cycles
CC  = CT * C
O1  = cycle_clock_offset router R1, interface towards R2
O2  = cycle_clock_offset router R2, output interface of interest
O1' = O1 + D
```

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 = ( ceil( ( O1' - O2 ) / CT) + C + 1) mod CC
map(i) = (i - 1 + A) mod 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) =~ 1.8, A will be 0, resulting in map(1) to be 1, map(2) to be 2 and 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 [Dmin...Dmax], 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(Dmax), and the controller plane needs to ensure that that A(Dmin)...A(Dmax) 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 [Dmin...Dmax] 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.

02

Refresh

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., Menth, M., and G. Cauchie, "Tree Engineering for Bit Index Explicit Replication (BIER-TE)", Work in Progress, Internet-Draft, draft-ietf-bier-te-arch-13, 25 April 2022, <https://www.ietf.org/archive/id/draft-ietf-bier-te-arch-13.txt>.

[I-D.ietf-detnet-bounded-latency]
Finn, N., Boudec, J. L., Mohammadpour, E., Zhang, J., and B. Varga, "DetNet Bounded Latency", Work in Progress, Internet-Draft, draft-ietf-detnet-bounded-latency-10, 8 April 2022, <https://www.ietf.org/archive/id/draft-ietf-detnet-bounded-latency-10.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