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

Requirements for Large-Scale Deterministic Networks

Abstract

Aiming at the large-scale deterministic network, this document
specifies the technical and operational requirements when the
different deterministic levels of applications co-exist and are
transported over a wide area.

Status of This Memo

Copyright Notice

Table of Contents

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. Various documents on data planes 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.

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 lower 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:

  |      <->| Industrial, tight jitter, hard latency limit
  |<------->| Industrial, hard latency limit
  |
  |<--------------.....>  Relatively lower latency requirements
  |
  |<-------------.....................>   Best effort
  |
  +-------------------------------------------------------->
                                                      latency
```
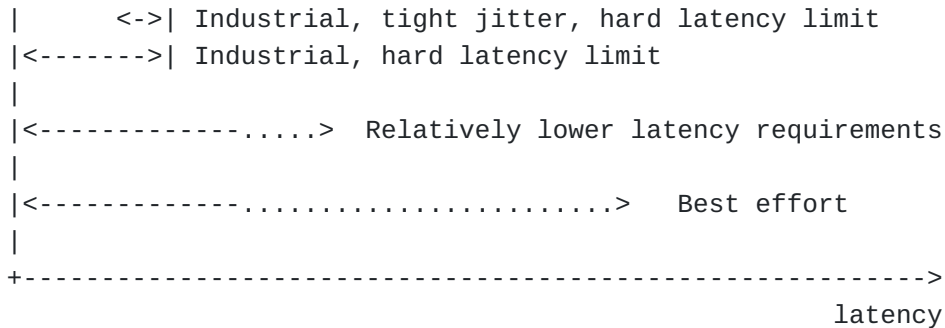
   Figure 1: Figure 1: Different levels of application requirements

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 administrators. 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.

```
+--------------+                          +--------------+
|              |      DetNet Connection   |              |
| TSN Domain I +--------------------------+ TSN Domain II|
|              |                          |              |
+--------------+                          +--------------+
              |       |        |        |        |
 Clock of TSN     +--------+--------+--------+--------+
 Domain I         =
                  =
                  =       |        |        |        |        |
 Clock of TSN     =      +--------+--------+--------+--------+
 Domain II        =      =
                  =<==D==>=
                  =      =
```
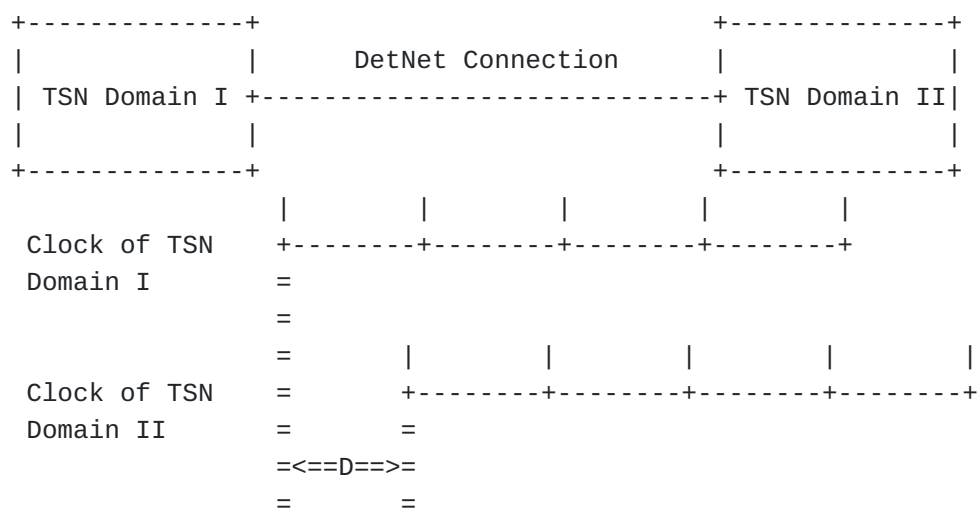
       Figure 2: Figure 2: Clock asynchrony between two TSN islands

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.

```
         Upstream Node X  |sending cycle  |             |
                          +--"-----------+-----------+
                          =  "\           =           =
                          =  " \          =           =
                          =  "  \         =           =
                          =  "   \        =           =
                          =  "    V       =           =
         Downstream Node Y |receiving cycle|          |
                          +--"----"-------+----\-------+
                          =  "    "       =     \      =
                          =  "    "       =      V resent out
                          =  "    "       =           =
             Time Line    -=--"----"-------=-------------=----->
             (in us)       0  |    |   10              20
                              v    v
                         Transmission Latency
```
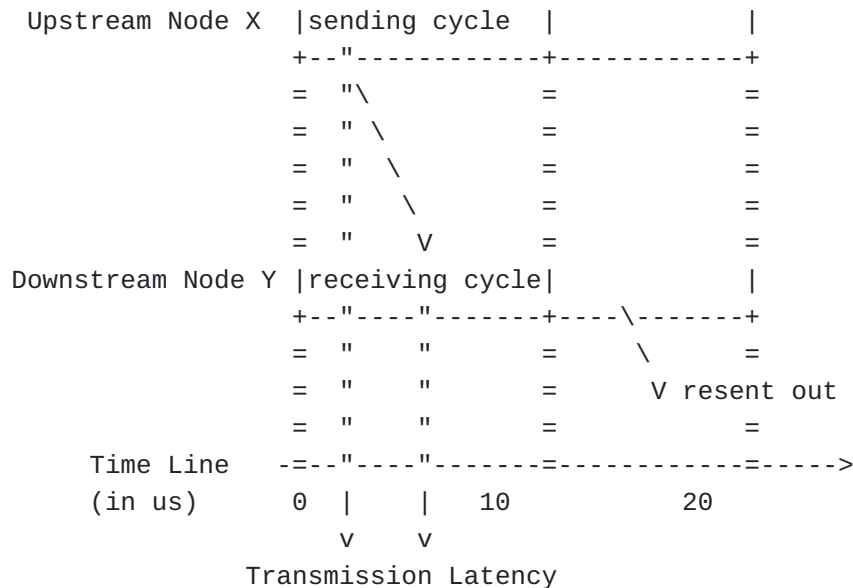
   Figure 3: Figure 3: The influence of transmission latency on a cyclic
                                method

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

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.

4.4.1. Be Scalable to Numerous Network Devices

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.

4.4.2. Be Scalable to Massive Traffic Flows

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. Conclusion

This document specifies the technical requirements when ensuring the
deterministic features in the large-scale networks. Some of the
proposed queueing mechanisms are analyzed 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.

6. Security Considerations

There are no IANA actions required by this document.

7. IANA Considerations

This section will be described later.

8. 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.

9. Contributors

The following people have substantially contributed to this document:

    Zongpeng Du
    China Mobile
    EMail: duzongpeng@chinamobile.com

10. 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>.

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