ETSI GR NGP 016 V1.1.1 (2019-11)

GROUP REPORT

Next Generation Protocols (NGP); Large-Scale Deterministic Network

Disclaimer

The present document has been produced and approved by the Next Generation Protocols (NGP) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference

DGR/NGP-0016

Keywords

framework, IP, jitter buffer

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

Copyright Notification

Contents

Intellectual Property Rights 4
Foreword 4
Modal verbs terminology 4
1 Scope 5
2 References 5
2.1 Normative references 5
2.2 Informative references 5
3 Definition of terms, symbols and abbreviations 5
3.1 Terms 5
3.2 Symbols 6
3.3 Abbreviations 6
4 Introduction 6
4.0 General 6
4.1 Motivations 8
4.2 Challenges & Requirements 8
4.2.0 General 8
4.2.1 Tolerance of a certain level of end-to-end jitter 9
4.2.2 Fast convergence as new services are created 9
4.2.3 Fine-grained and scalable resource reservation method 9
4.2.4 Tolerance of long link propagation delay 9
5 Framework of large-scale deterministic network 10
5.1 Overview 10
5.2 User-network interface 10
5.3 Resource-reservation signalling 11
5.4 Deterministic forwarding mechanisms 11
5.4.0 General 11
5.4.1 Scalable resource reservation at control plane 11
5.4.2 Scalable deterministic forwarding at forwarding plane 12
5.5 Auditing toolset 14
History 15

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Foreword

This Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) Next Generation Protocols (NGP).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The work item will describe a framework that enables a Layer 3 deterministic service over large-scale networks. Four functional components to construct the whole framework are:

1) User-Network Interface (UNI);

2) resource reservation signalling;

3) deterministic forwarding mechanisms; and

4) auditing toolset.

No specific technical solution will be recommended in this work item. However, some example mechanisms will be described in order to prove the effectiveness of the framework.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] IEC 61850: "Communication protocol manual".

NOTE: Available at https://www.naic.edu/~phil/hardware/sitePower/evd4/1MRK511242-UEN_-_en_Communication_protocol_manual__IEC_61850__650_series__IEC.pdf.

[i.2] IETF RFC 8557: "Deterministic Networking Problem Statement".

NOTE: Available at https://datatracker.ietf.org/doc/rfc8557/?include_text=1.

[i.3] IEEE 802$^{\text{TM}}$ Published TSN Standards.

NOTE: Available at https://1.ieee802.org/tsn/#Published_TSN_Standards.

[i.4] IEEE 802.1Qbv-2015$^{\text{TM}}$: "IEEE Standard for Local and metropolitan area networks -- Bridges and Bridged Networks - Amendment 25: Enhancements for Scheduled Traffic".

NOTE: Available at https://standards.ieee.org/standard/802_1Qbv-2015.html.

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

For the purposes of the present document, the following sympbols apply:

K	the size of aggregated resource reservation window
T	the length of a cycle

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

ABRW	Aggregated Bandwidth Reserved Window
DSCP	Differentiated Services Code Point
EXP	EXPerimental
IEC	International Electrotechnical Commission
IP	Internet Protocol
LAN	Local Area Network
MNO	Mobile Network Operator
MPLS	Multi-Protocol Label Switch
NP	Network Processor
PLC	Programmable Logical Controller
RTT	Round Trip Time
SDF	Scalable Deterministic Forwarding
SID	Segment IDentifier
SLA	Service Level Agreement
SRH	Segment Routing Header
SRR	Scalable Resource Reservation
TC	Traffic Class
TLV	Type/Length/Value
TSN	Time Sensitive Network
UDP	User Datagram Protocol
UNI	User-Network Interface
VR	Virtual Reality

4 Introduction

4.0 General

Deterministic IP aims at enhancing the current IP in order to provide deterministic services. Deterministic services here means deterministic latency, very low packet loss, as well as visualization of auditing. Deterministic latency here refers to bounded latency and bounded delay variance (i.e. jitter) [i.2].

Deterministic latency is different from low latency. The curve in Figure 1 shows the probability distribution of latency, low latency is to lower the upper bound of this curve, while deterministic latency tries to squeeze the curve to narrow it. In other words, deterministic latency aims to reduce jitter.

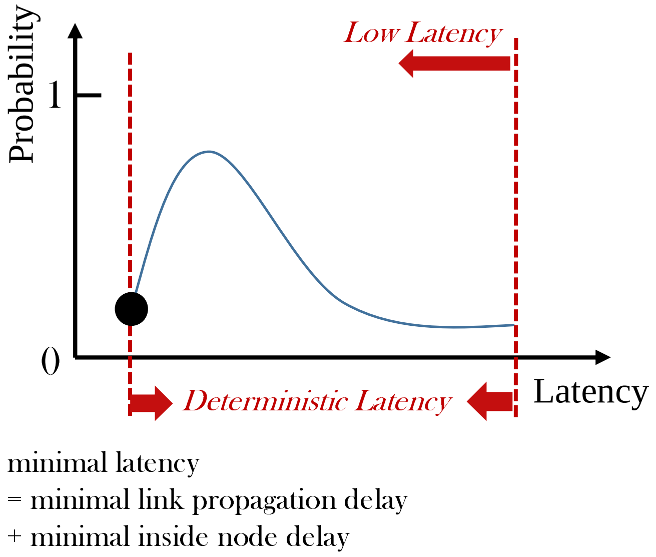


Figure 1: Low Latency vs. Deterministic Latency

Although deterministic latency and low latency have different definitions, their technical direction is not contrary to each other. Figure 2 analyses one-hop delay which mainly consists of two parts: on-link-delay and inside-node-delay:

- On-link-delay is mainly affected by the length of link, and the transmission rate of this link. Normally once the network devices have been deployed, their locations do not change. That means the length of link is usually constant. Meanwhile, the transmission rate of the link is dependent on cable media, almost invariable unless there is congestion on the link.

- Inside-node-delay refers to the time consuming intra-node operations, like queuing, NP processing, etc. Inside-node-delay varies a lot, and generates the long-tail effect as shown in Figure 1.

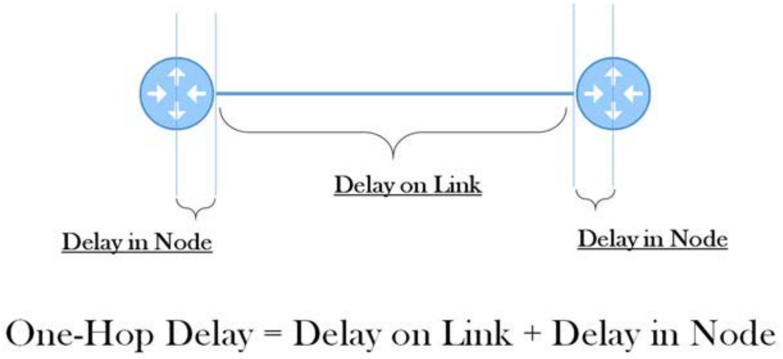


Figure 2: One-hop Delay Analysis

As Figure 3 shows, since on-link-delay is basically stable, the main way to reduce latency is to reduce the inside-node-delay. When the inside-node-delay is reduced to a certain level, it means the total delay variance is small, which is also deterministic latency. Hence, reducing the inside-node-delay can achieve both low latency and deterministic latency.

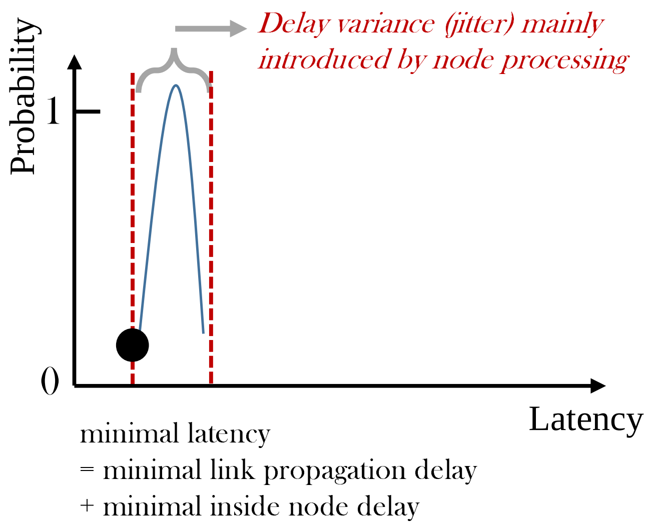


Figure 3: Reducing inside-node-delay to achieve both low latency and deterministic latency

4.1 Motivations

Deterministic latency is the premise that every node in a cooperative or scheduled system takes the right action at the right timing. There are several use cases such as stock exchange, relay protection of power network, online gaming, cloud PLC (Programmable Logical Controller), remote surgery, etc. The present document provides brief explanations to some of these use cases below to show why they require deterministic latency:

- Stock exchange: if a person/organization wants to buy a lot of shares, normally he will not buy all the shares at once. But through a complex program to monitor the price and generate a series of transactions, e.g. buy 1 share every 5 ms at a price of $1,5/share, or buy 1 share every 2 ms at a price of $1/share. If these transactions orders are transmitted over a non-deterministic network, then the rhythm of orders may be disrupted.

- Remote surgery: most surgeries require the cooperation of multiple doctors and nurses. If all of these doctors and nurses are operating remotely, the commands sent from different doctors (or nurses) should arrive at the patient in order, not too fast and not too slow. Deterministic latency here also could not be achieved through extremely low latency plus buffering due to many un-provisioning situations in surgery.

- Online gaming: for the sake of fairness, players' operational commands sent out at the same time, should arrive at the server and be processed at the same time. The "same time" here refers to the same time-slot, the faster the transmission and the less jitter a network provides, the finer a granularity time-slot can be divided by the server, so that smoother the operation the players will experience.

- Relay protection: two relay protection devices are placed at each end of power line, and send the same amount of current to the opposite end. Each relay protection compares its local current with the received current from opposite sides. If the diffirence of two currents is smaller than a threshold, then there is no error on line. Otherwise, something is wrong with the line. There is no time-synchronization among two relay protection devices, hence using half of RTT to estimate one-way latency. In order to ensure the replay protection system works correctly, one-way latency difference between two directions should be smaller than 200 µs, and their jitter should be smaller than 50 µs (IEC 61850 [i.1]).

4.2 Challenges & Requirements

4.2.0 General

Although there are several existing research on deterministic networks, none of them are able to provide deterministic forwarding over a wide area network with the following three features due to scalability:

- a large number of network devices;

- the distance between two network devices is long;
- a lot of deterministic flows over the network.

These above features will bring the following requirements:

- tolerance of a certain level of end-to-end jitter;
- fast convergence as new services are created;
- fine-grained and scalable resource reservation method;
- tolerance of long link propagation delay.

This subsequent clause will demonstrate these challenges in detail.

4.2.1 Tolerance of a certain level of end-to-end jitter

IEEE TSN [i.3] proposes a series of standards for determinacy in LAN. Most of these standards require for time synchronization among all devices within a LAN. TSN technologies are unable to apply directly to large-scale networks since these include a great amount of heterogeneous devices. Hence, it will be difficult and costly to keep precise time synchronization across all devices. The large-scale deterministic framework should be able to provide deterministic latency even under non-time synchronization scenarios.

4.2.2 Fast convergence as new services are created

In a local area network such as a factory, the information about when a deterministic service will start, how long the service will last, can be known in advance, or can even be planned. Based on this information, the local area network can adopt a global programming mechanism to calculate the accurate processing behaviours for each device, and achieve a global optimal performance. However, such global programming mechanisms like IEEE 802.1 Qbv [i.4] are unsuitable for service providers' networks. Many deterministic applications are expected to run on a service provider's network simultaneously. Different deterministic applications may have different lifecycles and SLA requirements, hence the network state changes dynamically. Such as VR communication may need to establish or tear-down the deterministic communication connections very frequently. If a mechanism relies on a stable network state for global computing, any change in network state (e.g. new application starts, or an application finishes, or SLA requirement changes) will lead to re-computing, even worse if all devices need to stop working and install the recomputed results, then this mechanism is hard to be deployed on service provider's network.

4.2.3 Fine-grained and scalable resource reservation method

In order to guarantee the QoS of deterministic flows, a network has to reserve necessary resources for each deterministic flow. All devices along the path should maintain a resource reservation state for an individual deterministic flow. The number of DetNet devices and flows in a network will be very dependent on the use case. 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 network may need to connect to ~100 K base stations (serving multiple MNOs'), and this number will only increase with 5G. If each ultra-low-latency slice or MNO is treated as a separate deterministic latency traffic flow (or tunnel), then even if each base station has a limited number of ultra-low latency slices or MNOs (e.g. ~10), there will still be a lot of, ~1 M, deterministic latency traffic flows on one network simultaneously. In such case, the per-flow resource reservation method is un-scalable.

4.2.4 Tolerance of long link propagation delay

IEEE 802.1 Qch [i.4] provides a typical and efficient cyclic forwarding mechanism that enables the end-to-end jitter be less than 2*T, where T is the length of a cycle. Figure 4 illustrates the methodology of IEEE 802.1 Qch [i.4]. Node A is the upstream node of Node B, the packets sent out by Node A at cycle x will be received by Node B at the same cycle. That is the length of a cycle which should be able to absorb the link propagation delay. Long link propagation delay can cause some troubles to IEEE 802.1 Qch [i.4]. In order to absorb the long link propagation delay, the length of cycle T needs to be set to a big value. However since packet's arrival time varies within the receiving cycle, larger cycle length means larger delay variance.

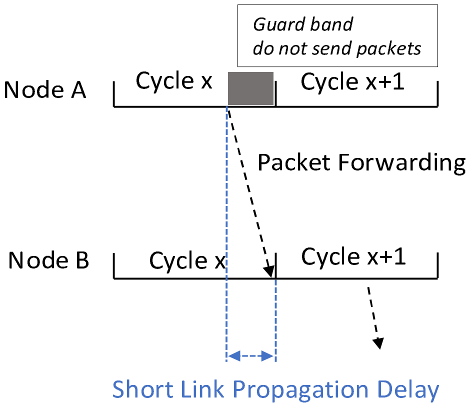


Figure 4: IEEE 802.1 Qch

5 Framework of large-scale deterministic network

5.1 Overview

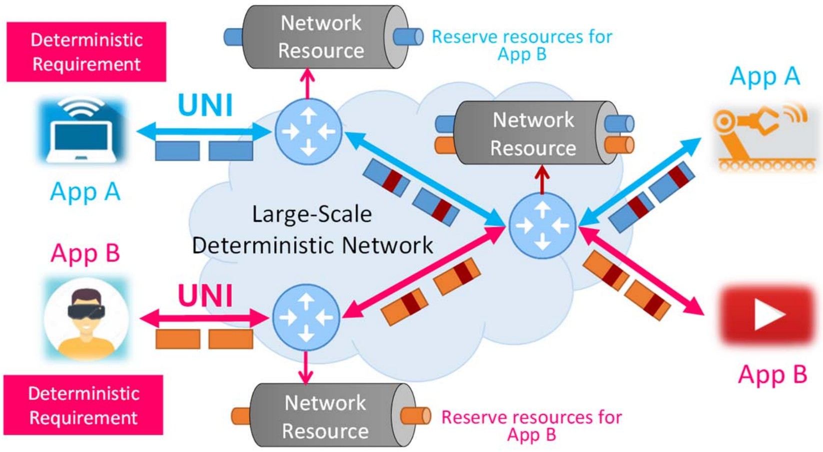

Deterministic Requirement
Network Resource
Reserve resources for App B
UNI
App A
Network Resource
App A
Large-Scale Deterministic Network
App B
UNI
App B
Deterministic Requirement
Network Resource
Reserve resources for App B

Figure 5: Functional components to enable large-scale determinacy

A high-level conceptual framework is shown in Figure 5. There are four necessary functional components:

1) User-Network Interface (UNI);

2) resource-reservation signalling;

3) deterministic forwarding mechanisms; and

4) auditing toolset.

5.2 User-network interface

Users express their deterministic requirements to network through UNI. There are at least two ways to implement the UNI:

- Through revising the protocol stack, let applications express the deterministic requirements by themselves.

- Through an agent device between user side and network side.

5.3 Resource-reservation signalling

After receiving the deterministic requirements from users, network should reserve necessary resources for these deterministic flows accordingly. There are three ways to do resource reservation:

- In-band signalling.
- Out-of-band signalling.
- Pre-configuration.

5.4 Deterministic forwarding mechanisms

5.4.0 General

Dedicated resource does not only deterministic forwarding, supporting forwarding plane and control plane mechanisms are also essential. Figure 6 shows a complete deterministic forwarding solution, include a Scalable Deterministic Forwarding (SDF) at forwarding plane and a Scalable Resource Reservation (SRR) at control plane.

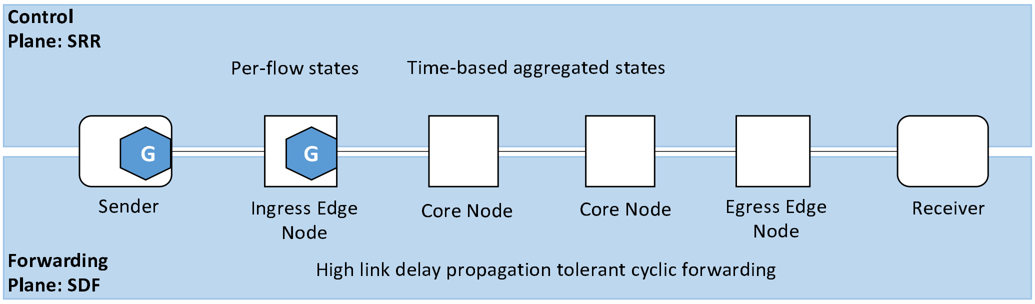


Figure 6: Deterministic forwarding architecture

In SRR, edge node records the resource reservation status for each individual flow and a core node aggregates the per-flow statuses. At the forwarding plane, all nodes are frequency synchronized, and ingress edge nodes or senders have a function called "*gate*" to shape the traffic flows into a certain pattern. Packets are forwarded in a cyclic queueing and scheduling fashion.

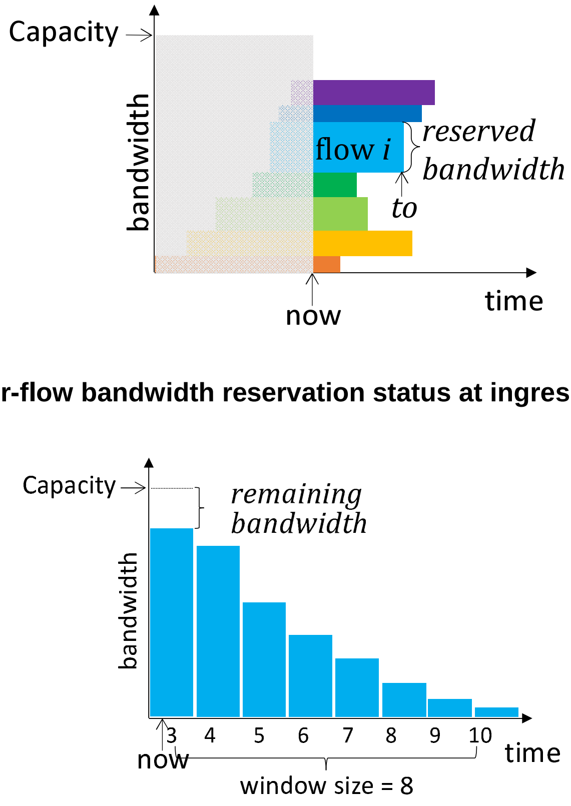
5.4.1 Scalable resource reservation at control plane

The Control plane should record the resource reservation status for each individual flow including which resources (e.g. an amount of bandwidth) are reserved for which duration. Such per-flow status contains at least (flow_ID, reserved_bandwidth, start_time, end_time) information as shown in Figure 7 (a). Maintaining per-flow status may be acceptable to edge nodes, since the number of flows at each edge will not be too high. However as a great amount of flows converge at core nodes, code nodes get overloaded, so the system will not be scalable unless aggregation is supported at core nodes. Figure 7 (b) illustrates the aggregation method in SRR (Scalable Resource Reservation):

1) Dividing time into time slots, then the per-flow status becomes:
   - flow_ID;
   - reserved_bandwidth;
   - start_time_slot;
   - num_time_slot accordingly.
2) The core node calculates the sum of reserved_bandwidth in a time_slot. For a time slot, core node just needs to maintain a variable. Supposing that a core node can maintain K time slots' statuses, i.e. the size of Aggregated Bandwidth Reserved Window (ABRW) is K.

3) A new resource reservation request succeeds only if there are sufficient resource along the path. Resource should be reserved in an integer of time slot. This is even if a flow just needs to use 1,7 time slots' resource, core node still needs to reserve 2 time slots' resource for it. Furthermore, a flow can request at most K time slots every time. If more than K time slots are needed, a sender should send a renewal request before the expiration of K time slots.

4) The Core node refreshes its ABRW according to the per-flow status maintained at edge nodes. The sender also can actively tear down its resource reservation.



(a) Per-flow bandwidth reservation status at ingress node

(b) Aggregated bandwidth reserved window at core node

Figure 7: Scalable resource reservation

5.4.2 Scalable deterministic forwarding at forwarding plane

Cyclic forwarding is unaware of per-flow state and can guarantee deterministic latency as illustrated in Figure 8. In cyclic forwarding, time is divided into several cycles, and packets are sent cyclically. The cycle number of sender and the cycle number of receiver are fixed, while the exact timing of cycles are unknown. Hence the best situation is packets sent out at the end of sending cycle and received at the beginning of receiving cycle. Without considering packet loss, the worst situation is if packets are sent out at the beginning of sending cycle and received at the end of the receiving cycle. Therefore, the end-to-end jitter's upper bound = "end-to-end delay of worst case" - "end-to-end delay of best case" = 2*T.

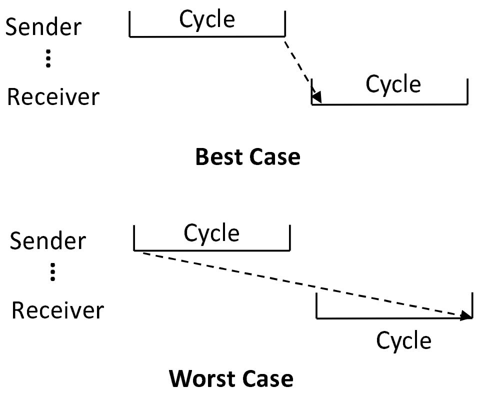


Figure 8: Cyclic forwarding

As clause 4.2.4 points out, TSN-Qch is a classical cyclic forwarding mechanism. However it requires time synchronization among all nodes, and limited link propagation delay. Both of these requirements are inappropriate to large-scale network. SDF inherits cyclic forwarding's advantage while breaking the limitation that TSN-Qch has. In SDF, all nodes are frequency synchronized as Figure 9 illustrated, and long link propagation delay is tolerated. Meanwhile, the end-to-end queuing delay is 2*T*hops.

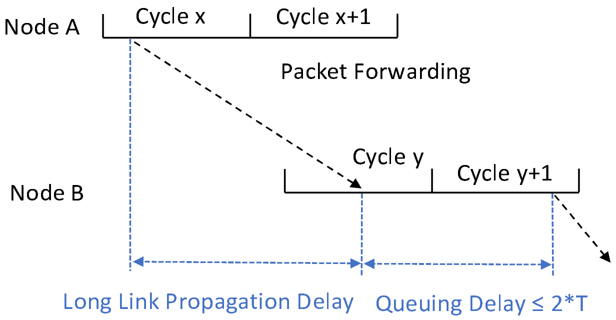


Figure 9: Scalable deterministic forwarding (SDF)

Without considering the non-queuing propagation delay variation, each pair of neighbouring nodes has a stable cycle mapping relationship, that could be used to indicate the packet forwarding time. As an example shown in Figure 10, Node A is the upstream node of Node B, the cycle mapping relationship between A and B can be expressed as cycle x→ cycle y+1. That is packets sent out by Node A at cycle x will be re-sent out by Node B at cycle y+1. Since there is no time-synchronization between them, Node B may receive two cycles' packets as the example shows. Hence each output port requires 2 receiving queues, and 1 sending queue. Accordingly, each packet needs 2 bits to carry cycle-identifier in order to indicate while queue it should enter.

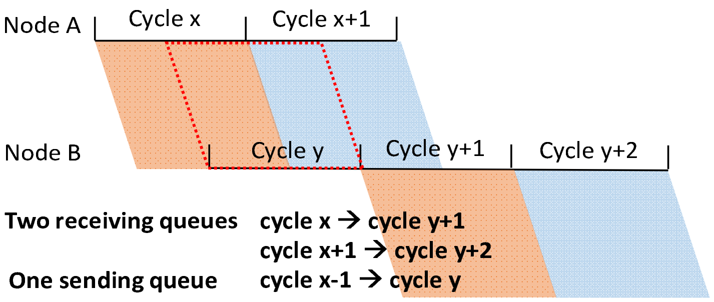


Figure 10: Three queues in SDF

There are several ways to carry this 2 bits cycle identifier, an incomplete list:

- DSCP of IPv4 Header.
- Traffic Class of IPv6 Header.
- TC of MPLS Header (used to be EXP).
- IPv6 Extension Header.
- UDP Option.
- SID of SRv6.
- Reserved Field of SRH.
- TLV of SRv6.
- TC of SR-MPLS Header (used to be EXP).
- 3 (or 4) labels/adjacency SIDs for SR-MPLS.

5.5 Auditing toolset

A fine-grained auditing toolset should be provided for users, so that users can verify if their requested deterministic services are really offered. Moreover, such fine-grained auditing toolsets should not introduce any scalability problems. Therefore, the in-band approach may be the most appropriate one.

History

Document history		
V1.1.1	November 2019	Publication