INTERNET-DRAFT
Intended Status: Standard Track
Expires: May 4, 2017

Y.Yu
HUAWEI Technologies
October 31, 2016

Layer 3 Quantized Congestion Notification(L3QCN)in the Converged Network
draft-yu-tsvwg-l3qcn-00

Abstract

The more demands for the lossless and low latency network in the modern datacenter appear because the proliferation of demanding applications. Some congestion control schemes such as CN, PFC, ETS which is introduced by IEEE 802.1 focus on the L2 network domain. While current TCP/IP stacks can’t meet these requirement on L3 or above networks. This draft introduces the L3QCN(Layer 3 Quantized Congestion Notification), an end to end congestion control scheme which adopt QCN and DCQCN on L2 network. It specifies protocols, procedures, and managed objects to support congestion control on the datacenter network.

Status of this Memo

This Internet-Draft is submitted to IETF 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 http://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 May 4, 2017.

Copyright and License Notice

Table of Contents

1 Introduction

Currently, there are 3 classes of streams in the DC network:
1)Storage Traffic (Lossless)
2)High Compute Traffic(Low latency)
3)Ethernet Traffic (Certain packet loss& latency tolerance)
Traditional DC network treat different traffic with different network bearer which exist in the small scale DC. While with the expand of the DC scale, there is an available method which use the Ethernet to bear the streams by applying the congestion control method. IEEE has introduced the following specifications:

1. Enhanced Transmission Selection (ETS) [1] When the offered load in a traffic class doesn’t use its allocated bandwidth, enhanced transmission selection will allow other traffic classes to use the available bandwidth. This avoid the burst of one class traffic to influence other classes which provide the minimum guaranteed bandwidth to all traffic classes. This also facilitate the multiple classes exist in one network.

2.Priority-based Flow Control(PFC) [2] Data Center Bridging networks (bridges and end nodes) are characterized by limited bandwidth-delay product and limited hop-count. Traffic class is identified by the VLAN tag priority values. Priority-based flow control is intended to eliminate frame loss due to congestion. This realized the lossless of storage stream and no impact to other 2 traffic classes when all the 3 traffic classes coexist in the Ethernet.

3.Quantized Congestion Notification (QCN) [3] This mechanism enable bridges to signal congestion information to end stations capable of transmission rate limiting to avoid frame loss. Resolve the latency increase caused by flow control or packet retransmission to achieve the higher network throughput.

This draft introduce a L3QCN method to resolve the congestion problem under the converged network in the datacenter. Different classes of traffic will be configured with corresponding priorities. Bridge will apply the policies of congestion control according to the traffic of congested traffic which is defined by the priority.

1.1 Terminology

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].

2 Current Congestion Control method

2.1 QCN Introduction

2.1.1 QCN Technical Solution

QCN is defined in IEEE 802.1Qau, there are 2 types of Ethernet frame: One data frame with CN-TAG as Figure 1 shown. The Converged Network Adapters (CNA) which support QCN function will send out the CN-TAG frame when connecting network domain. The difference from the normal frame is the CN-TAG field in the head of Ethernet frame which includes RPID (also known as FLOW-ID). RPID will uniquely identifies every stream sent by the adaptor. When the congestion appeared, bridge will send out CNM frame(introduced in second clause) to notify the source node to stop sending this stream. The FLOW-ID of source frame will be encapsulated in the CNM frame. When the adaptor receives the CNM frame, it will reduce the transmission rate of the identified flow in order to control the specific traffic precisely.

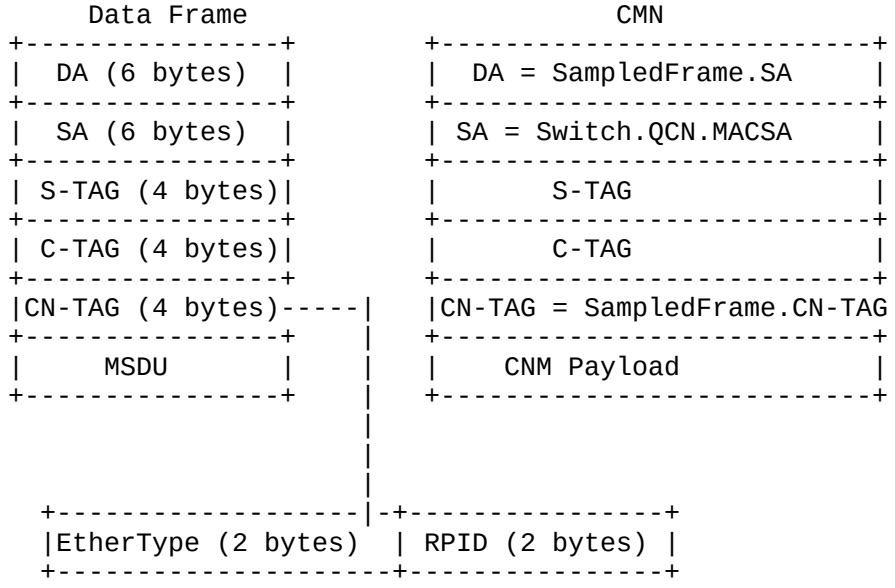


Figure 1. QCN data frame and CMN

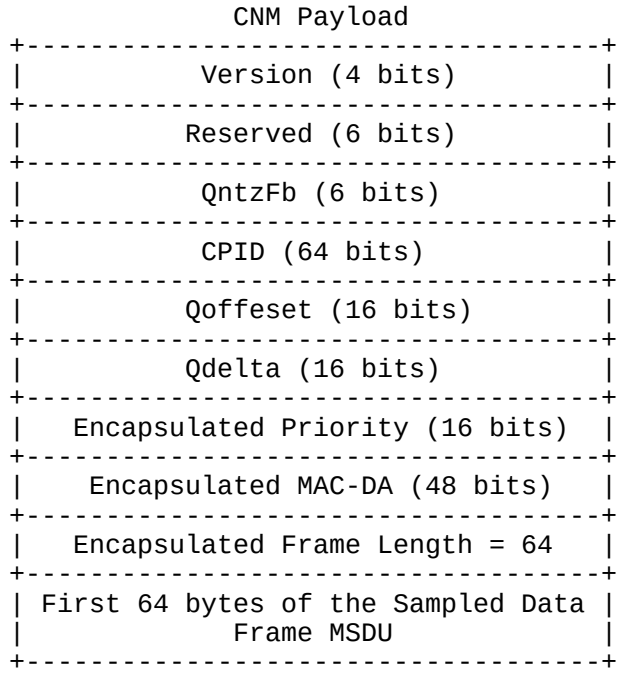


Figure 2. CMN payload

The CNM frame is shown as Figure 2:

Field 1: Version of CNM message (4 bits)
Field 2: Reserved (6 bits)
Field 3: QntzFB, Quantized feedback of CNM message (6 bits)
Field 4: Congestion Point Identifier (CPID, 8 bytes). In order to assure the uniqueness of the identifier, use the MAC address as the upper 6 bytes. Lower 2 bytes identify the different ports or different priority classes in the same device.
Field 5: QOffset (2 bytes). Current number of available bytes in the sending queue of the congested point (CP)
Field 6: QDelta (2 bytes), the difference of available bytes of CP at 2 time point.
Field 7: Encapsulated priority (2 bytes). Use upper 3 bits of the 1st byte to fill the priority of the CNM frame. Else is 0.
Field 8: Encapsulated destination MAC address (6 bytes). Fill the destination MAC address which trigger the CNM frame.
Field 9: Encapsulated MSDU length (2 bytes). The length of the Encapsulated MSDU.
Field 10: Encapsulated MSDU (64 bytes). Fill in the payload of the CNM.

2.1.2 The limitation of QCN

During the congestion, bridge need to encapsulate the FLOW-ID(in the head of Ethernet frame) in the CNM. Then replace the destination MAC address of CNM with the source MAC address of the congested frame in order to ensure CNM could be send back the sending server. Sending server reduce the flow according to the FLOW-ID carried in the CNM. This characteristic limit QCN only in Ethernet(Level 2 in ISO). Since the head of Ethernet frame will be changed during every packet routing in the IP network, the FLOW-ID and MAC address of sending server will be lost. So the downstream bridge could not create the CNM and send back to the sending server. QCN couldn't support the Layer 3 networking.

2.2 Introduction of DCQCN
2.2.1 DCQCN technical solution

DCQCN[4] is a kind of congestion control solution proposed by Microsoft for the DC network domain. DCQCN is mainly deployed in the RoCEv2 scene. CP (Congestion Point, bridge) set the CN(congestion notification) for the datagram with probability according to the degree of the congestion. After the datagram sent to NP(Notification Point, receiving server) , NP construct CNP (Congestion Notification Packet) to RP(Reaction Point, sending server). RP reduce or increase the transmission rate according to the dedicated algorithm which is similar to QCN.

2.2.2 The limitation of DCQCN

DCN construct ECN(explicit congestion notification)[5] tag during the congestion and forward to NP. NP construct CNP to notify RP. The reaction is not quite timely(Control Loop Delay is big). If the congestion appeared on the upper jump, for example on the TOR, there is more delay of 9 jumps than the direct response.

3. Layer3 QCN
3.1 L3QCN Introduction

L3QCN is a technical solution to resolve the congestion problem under the converged network in the datacenter. Different class of traffic will be configured with corresponding priorities. Bridge will deploy the policies of congestion control according to the class of congested traffic which is defined by the priority.

3.2 Use case of L3QCN
3.2.1 A hybrid method with QCN
Deploy priority 5 to the traffic sent out by QCN server. When the queue buffer for the priority 5 exceed the defined threshold, the bridge will back-haul the congestion information to the accessing TOR. TOR and HOST can reach to each other on the Ethernet which is similar to a L2 domain. In this situation, standard QCN is performed. Accessing TOR transform the congestion information to standard CNM frame and send to QCN server

which realize the congestion control.

Deploy priority 7 to the traffic sent out by RoCEv2 server. When the queue buffer for the priority 7 exceed the defined threshold, CP judges the key flow causing the congestion. Then CP construct the standard CNP. The RoCEv2 server reduce the transform rate according to the probability of CNP reception.

3.2.2 L3QCN in CLOS fat-tree

L3QCN control steps are as follows:

1)Datagram sent out from QCN server enters the accessing TOR. Firstly, accessing TOR will save the source MAC address, FLOW-ID, VLAN-TAG and IP 5-tuple to the local table, shown in Table 1. Then TOR perform the normal routing.

Src IP	Dst IP	Src Port	Dst Port	Proto -col	MAC SA	Flow ID	VLAN TAG
192.168.2.100	192.168.3.30	5678	21	6	0x01a4f5aefe	0xa878	100
10.1.10.2	10.2.20.1	8957	21	6	0xfd16783acd	0xc9a0	1024
192.168.2.100	10.3.50.1	2345	80	6	0x0a25364101	0x0ac9	3
200.1.2.3	100.2.3.4	2567	47	17	0xed16d8ea0a	0x37a0	90

Table 1. FLOW-ID Mapping Table

2)Shown in Figure 3. Congestion caused by Incast flow, T4 detect the congestion in a certain queue and exceed the threshold. Distinguish the flow model according to the priority of the queue.

Please view in a fixed-width font such as Courier.

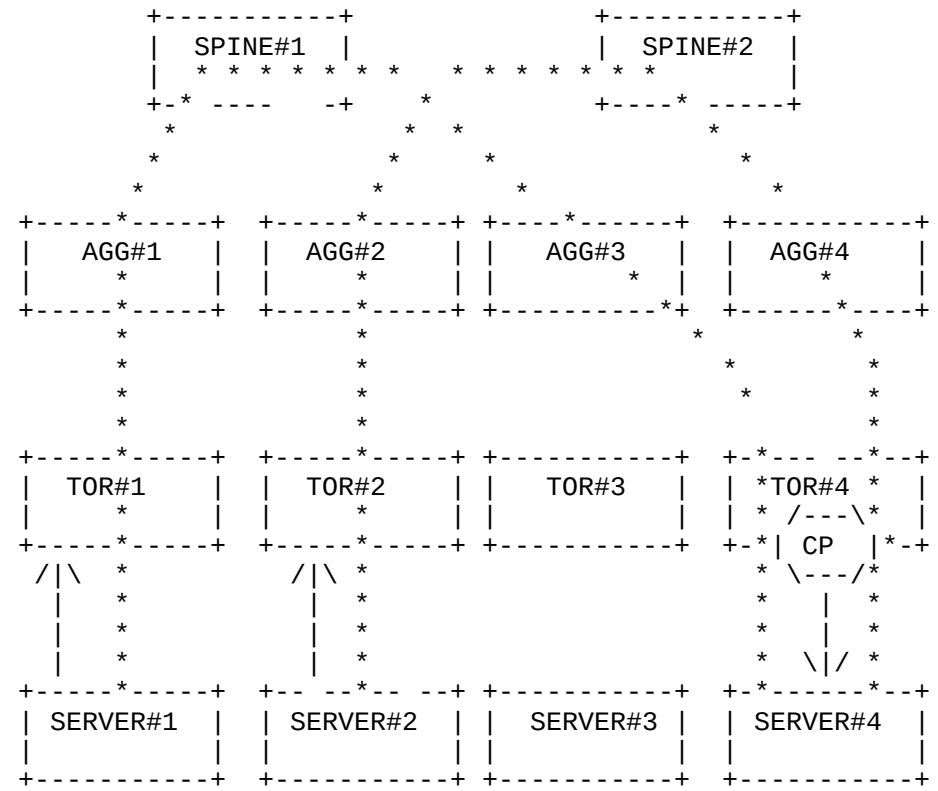


Figure 3. Incast flow model

3)If it is the flow from QCN server, conduct self-defined CNM which include the 5-tuple, congestion indications (defined in QCN specification, such as QntzFb, CPID, Qoffset,QDelta), encapsulate IP +UDP. UDP need to use a specific port No. which is used to recognize the QCN frame in TOR. Or use a bit in the IP head(reserved bit) to indicate the type of the frame. The dedicate IP is set to the source IP which assure the CNM could be routed to the accessing TOR. It’s better to construct the self-defined CNM based on the standard CNM to reduce the writing times which might increase the performance.

4)As shown in the Figure 4&5 , T2 recognize the self-defined CNM according to the destination UDP port. T2 map the self-defined CNM to the standard CNM and send to H2. The QCN is performed in L2 domain because the adaptor of H2 support the standard QCN.

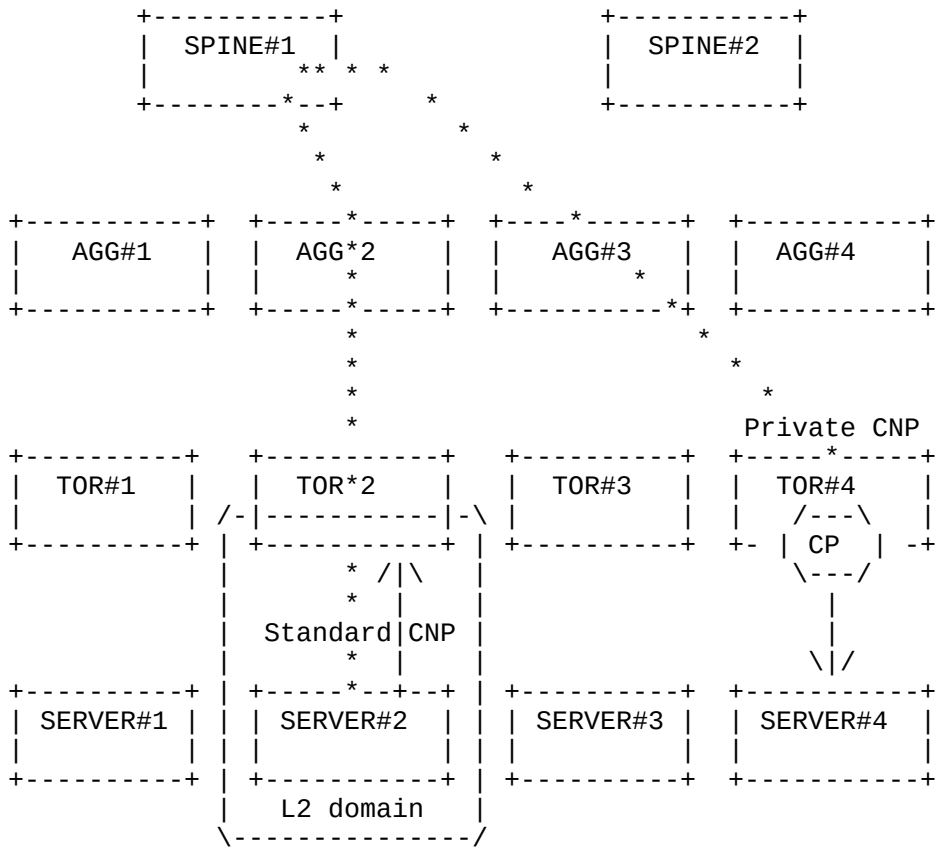
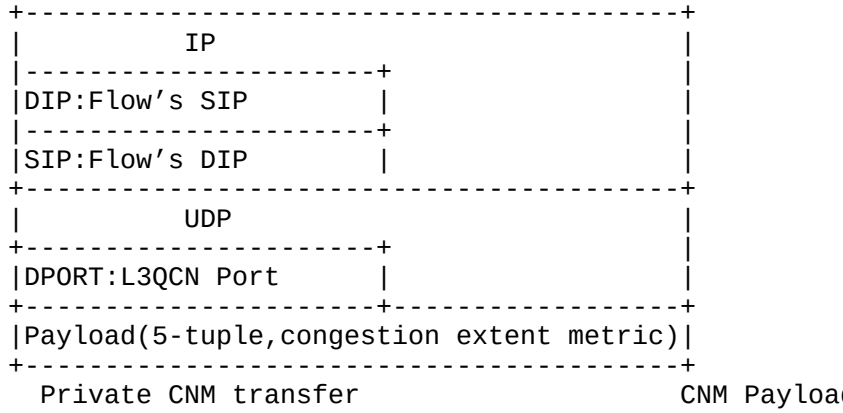


Figure 4. Construct the self-defined CNM

Please view in a fixed-width font such as Courier.



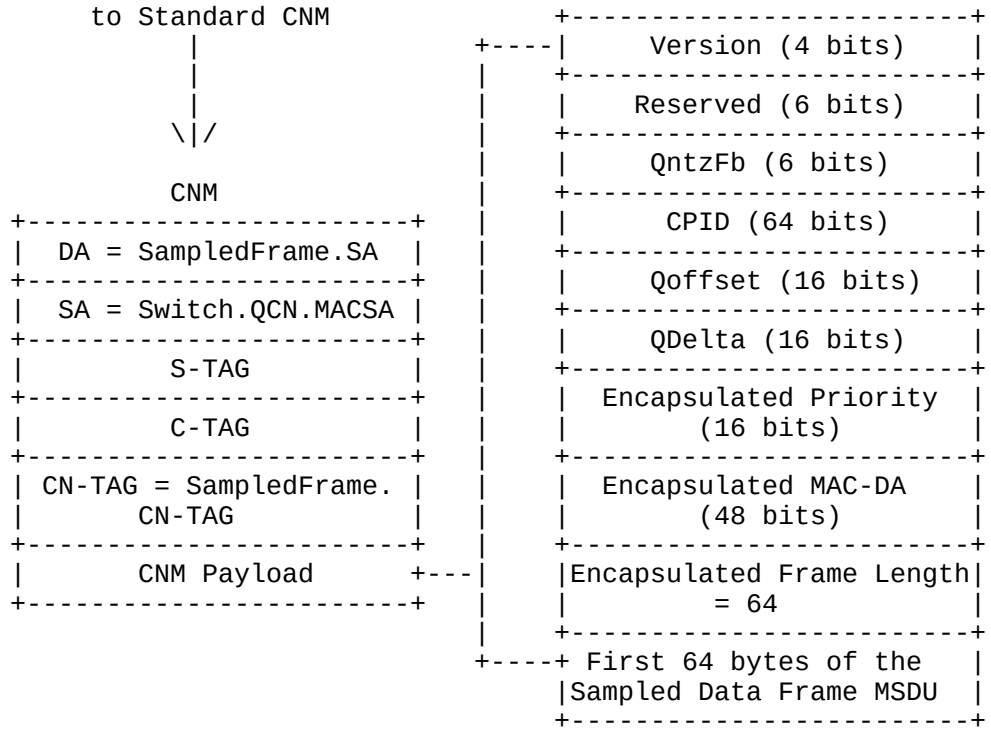
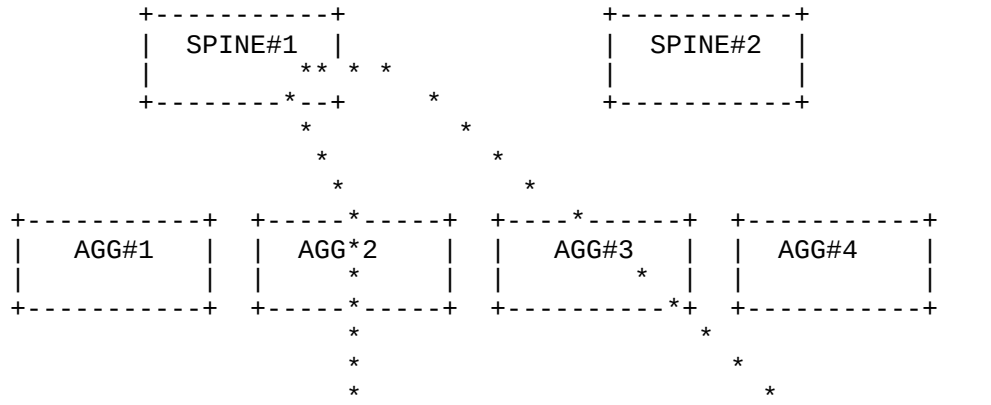


Figure 5. Transfer Private CNM to Standard CNM

5)As shown in the Figure 6 , T4 recognize which flow causes the congestion. CP construct the standard CNP. The adaptor of RoCEv2 server support CNP and reduce the transmission rate according to the probability of CNP reception.

Please view in a fixed-width font such as Courier.



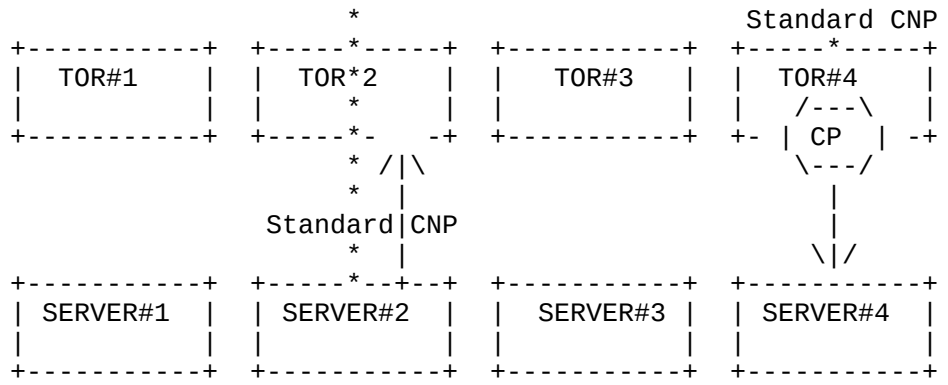


Figure 6. CP construct the standard CNP based on RoCEv2

4. Conclusion

L3QCN resolve the problem that QCN could not support L3 network. L3QCN realize the QCN mechanism across the L3 network. There is no modification on the QCN servers.
For the RoCEv2 traffic, since the CP send the CNP when reach the congestion threshold, it reduce the Control Loop Delay dramatically which could reduce the depth of the queue buffer and the datagram delay. The performance of the network is improved.

5 Security Considerations

N/A

6 IANA Considerations

Will apply the specific UDP port No. if required.

7 References

7.1 Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

7.2 Informative References

[1] IEEE 802.1: 802.1Qaz Draft 2.5- Enhanced Transmission Selection

[2] IEEE 802.1: 802.1Qbb Draft 2.3- Priority-based Flow Control

[3] IEEE 802.1: 802.1Qau Draft 2.4- Congestion Notification

[4] Yibo Zhu et al., SIGCOMM 2015, Congestion Control for Large-Scale RDMA Deployments

[5] K. Ramakrishnan, S. Floyd, and D. Black. The addition of explicit congestion notification (ECN). RFC 3168

Authors' Addresses

Yolanda Yu
101 SOFTWARE AV., YUHUATAI DIST., NANJING,
JIANGSU,210012,CHINA
EMail: yolanda.yu@huawei.com