Multicast Congestion Control issues with 25G/50G/100G Ethernet

christoph@lameter.com

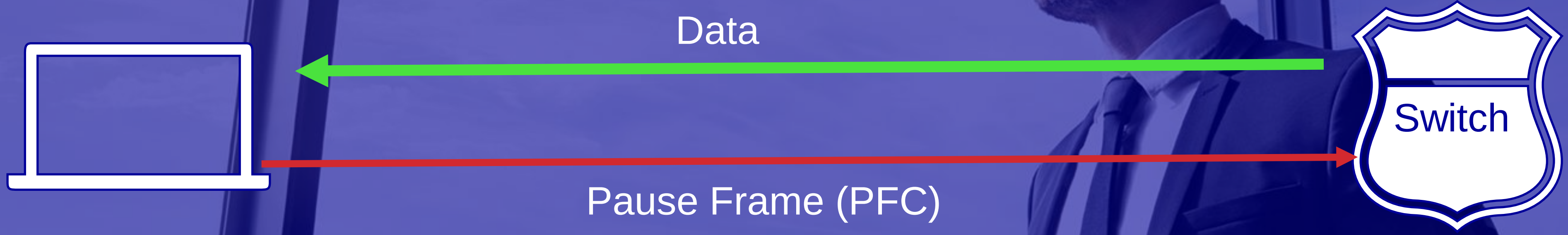
March 27th, 2023

DEUTSCHE BÖRSE GROUP

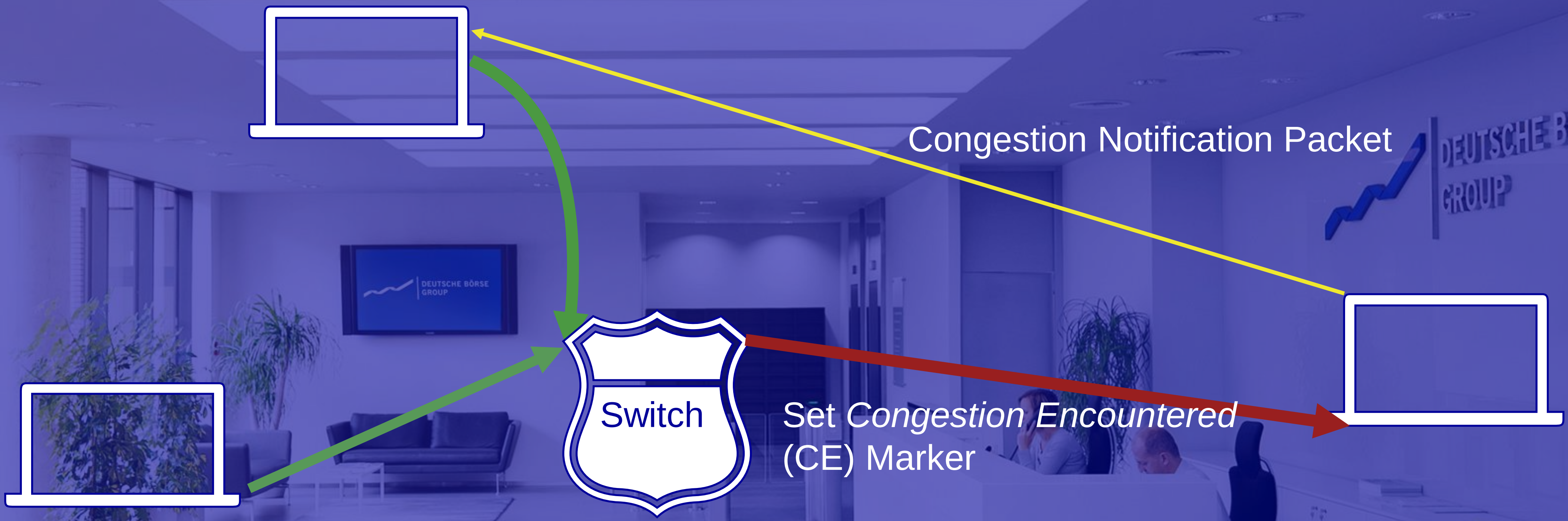
The problem in 25G/50G/100G networks

* Megabytes of packet data are transmitted in the period of a round trip time involving a host processor or firmware.
* Packet loss may cause massive retransmits and/or long stalls of a data stream.
* Measures were introduced to avoid packet loss by moderating the speed of the senders.
* TCP was modified to support ECN which allows the control of the speed of packets coming from specific ports on a system (RFC3168)
* Data Center Bridging Extension (IEEE802.1Qbb) to implement flow control for Ethernet links.

Basic Ethernet Flow Control (IEEE802.1Qbb)



RFC3168 / ECN: Per Flow Congestion Control



Non TCP Example: IBTA Spec A17: Roce V2 Congestion Control

ECN / CE Signalling

A17.3.1.1.3 / 17.9.3 specifies the use of the ECN field to indicate congestion in the switch to the receiver. The receiver can then send CNP packets to slow down the sender. Support for ECN / CE is optional (CA17-5).

CNP packets slow down senders

A receiver sends CNPs to the individual sender if the CE flag is set in the header (CA17-45, CA17-44). Switches have advanced strategies to mark packet more or less frequently depending on the level of congestion to trigger CNPs.

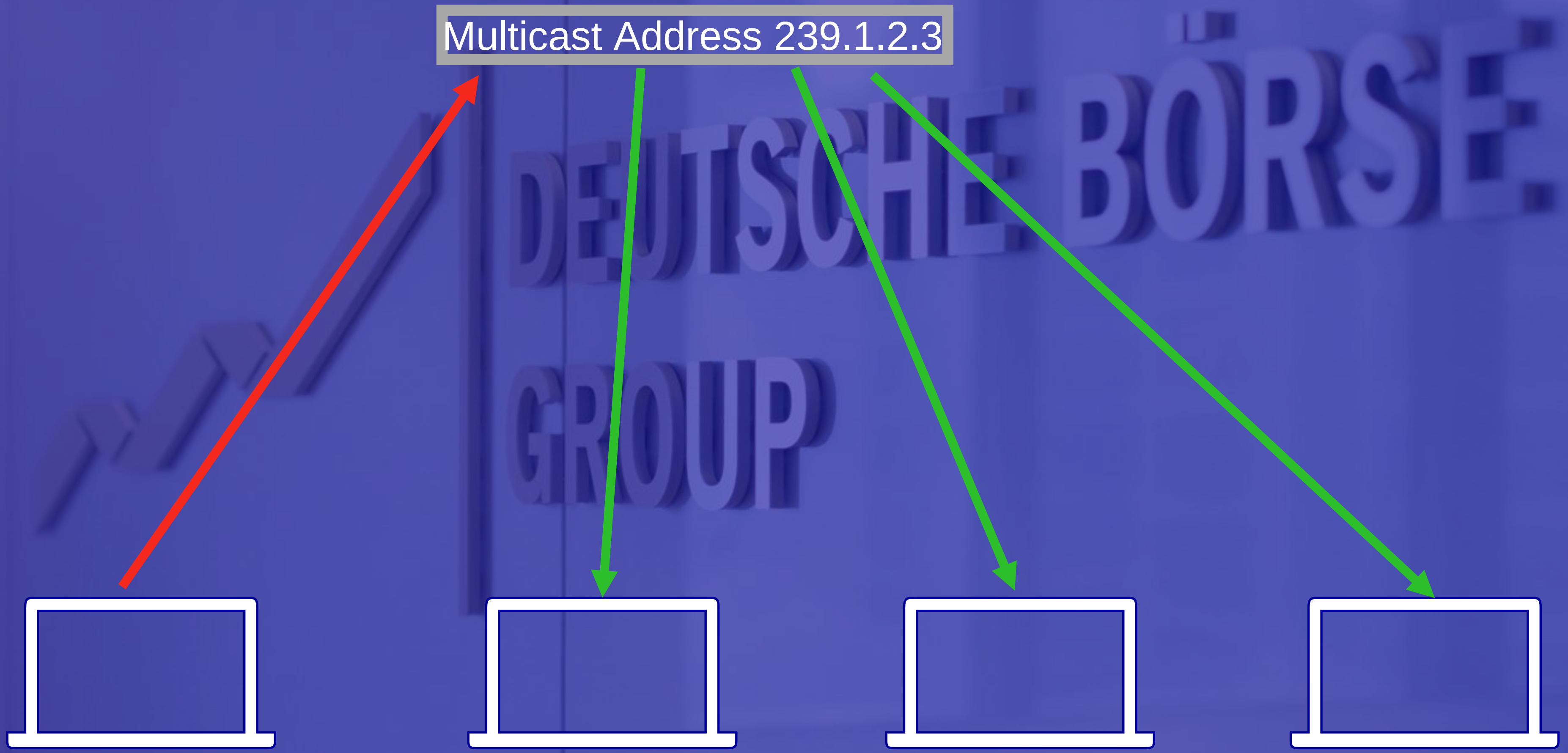
Link-Layer Flow-Control IEEE802.1Qbb

Roce may also use Link Layer flow control (Pause Frames, PFC) in addition to CNP packets. This is a defined flow control method at the *Ethernet Layer* that ROCE packets can make use of. The sending rate on the one side of a cable is slowed down to mitigate network congestion.

Lossless operations A17.9.1.
ROCE handling should be lossless. This is in practice realized in the following way:

1. CNP packets to slow down the sender
2. Activation of Link Layer Flow control if CNPs are not sufficient.

Multicast : Magic in Networking

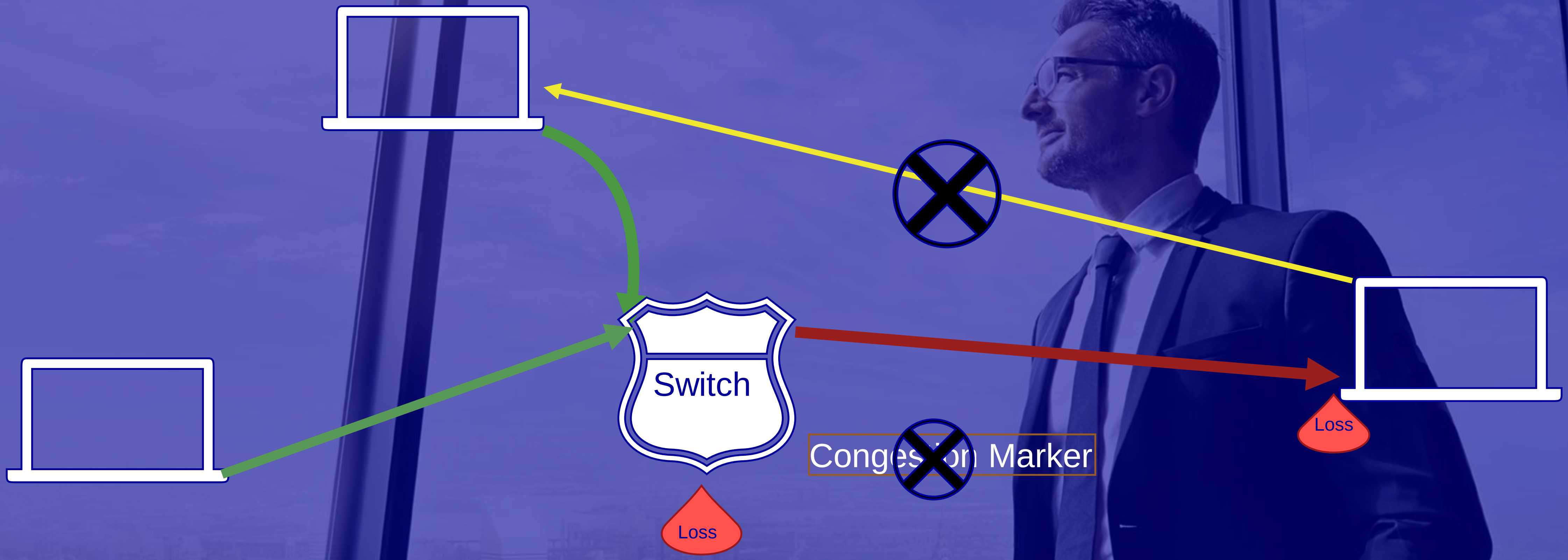


BUM : The second class of Ethernet packet

A technical Ethernet term referring to a packet without a single well-known destination

Broadcast
Unknown Unicast
Multicast

Current Handling of Multicast Congestion



Known Ethernet Switch support for Congestion Management

Feature	Unicast PFC	Unicast CNP	Multicast PFC	Multicast CNP
Cisco	Yes	Yes	Potentially possible with some modification in current ASICs	Cisco switches may be able to do this. Under investigation by Cisco
Nvidia	Yes	Yes	Yes	Vendor Project
Arista	TBD	TBD	“Should work” (David Snowdon)	TBD

Known RDMA NIC Support for ROCE v2 Multicast

Feature	Receive	Loopback Send	Remote Send
Nvidia	Yes	Yes	Yes
Intel	Yes	Yes	Patch for testing against upstream drivers exist to enable multicast send
Broadcom	No MC support in Linux source tree	-	-

Current handling of Roce V2 multicast traffic

- Link Level flow control is possible if the switch supports in (one known vendor). PFC frames will slow down all traffic on the link. Otherwise, traffic will be dropped.
- CE bit is not set by the switches if links become congested since switches do not support congestion control for Multicast.
- Therefore, the NICs do not send CNPs.
- Traffic on the link does not slow down.
- Massive packet loss is possible which leads to unrecoverable failure of applications expecting reliable multicast solutions as provided by middleware vendors.
- Multicast based RDMA applications work reliably on Infiniband but not on ROCE.
- This is contrary of A17.9.1 which states that ROCE v2 handling should be lossless. However, that statement seems to be only be true for unicast traffic. Annex 17 does not make a distinction between multicast and unicast and does not discuss the issues related to multicast congestion control.
- Apart from one vendors support for Multicast Flow Control, currently available switches have no option of making ROCE multicast reliable. The one vendor does not support CNPs so all traffic coming from an interface must be throttled instead of only the stream that causes the overload.

IETF Handling of ECN through CE

DEUTSCHE BÖRSE GROUP

RFC 3168 is primarily concerned with defining a congestion control mechanism for the TCP protocol by allowing detection of congestion before queues overflow. ECN is relying on the transport protocol (Section 6) to allow a reaction to packets that have encountered congestion. The receiver notices that the CE bits are set and performs a transport specific reaction to reduce congestion in the fabric.

RFC 3168 envisions other transport protocols to be using this mechanism mentioning among others "*unreliable and reliable multicast transport protocols*" (Section 6).

ROCE V2 is such a transport protocol where the IBTA has defined congestion control mechanism through the congestion notifications packets (CNP). These have been so far be seen as restricted to unicast. It looks like it is within the intended scope of RFC 3168 if the use of CNP would be extended to cover multicast as well.

The IBTA spec does not restrict the use of CNPs to unicast. However, the actual implementations by the hardware vendors have so far restricted the use of CNPs to unicast.

CNP replies as a result of receiving Multicast Packets marked with CE

A17.9.3 states:

CA17-44: If RoCEv2 Congestion Management is supported, upon receiving a valid RoCEv2 packet with a value of ’11 in its IP.ECN field the HCA shall generate a RoCEv2 CNP formatted as shown in Figure 360 on page 2000 directed to the source of the received packet. The HCA may choose to send a single CNP for multiple such ECN marked packets on a given QP.

So far, the CNPs are only sent as a result of receiving unicast ROCE v2 packets. The use of this mechanism when receiving a multicast packet is straightforward. A unicast CNP packet is sent when a multicast packet with the CE bits set is detected. The destination is the sender of the multicast traffic which is the source address of the multicast packet (which is a unicast address).

The sender will then receive the unicast CNP packet and moderate the output of the multicast QP to reduce congestion on the fabric.

Multicast and IEEE 802.1Qbb (Flow Control via PFC / Pause)

IEEE 802.1Qbb considers any traffic between the two endpoints of a link. The standard is therefore not concerned with unicast versus multicast. The PFC frame itself is a multicast frame through. However, that special frame is only used for communication between the two endpoints of the link. The PFC frame is never forwarded to other ports of the switch. The standard establishes priority classes for PFC support (defined in IEEE 802.1Q) and allows the slowing down of senders for packets of a particular category.

This means that if unicast and multicast packets are treated the same then the mechanism works for either type of packet.

This is the case for switches from Nvidia. However, Cisco switches do not apply priorities to multicast traffic. Cisco has built a special queue for multicast traffic in which all multicast traffic accumulates bypassing the queues for the traffic classes. There are no provisions for congestion support for this special queue. This then defeats the mechanism envisioned in IEEE 802.1Qbb. Cisco switches not send PFC frames if multicast traffic is congested.

RDMA NICs maybe configured send PFCs to the switch if the applications are slow in picking up the traffic. This does work with the Cisco switches resulting in traffic to accumulate on the switches. The Cisco switch will have to drop packets when too many packets accumulate since it is not able to send PFCs to the sender of the multicast traffic.

CNP Flooding concerns

Multicast replication can result in single packets marked with the CE bit arriving at multiple endpoints which may all send CNP responses to the sender. It is therefore advisable to be able to limit the number of CNPs in a certain time interval. This has already been an issue for unicast and therefore these throttling mechanisms already exist at the sender/receiver as well as the switches in the hardware known to the author. These are currently not used for multicast. There are 3 points the network where these flooding issues could be handled:

1. The receiver may receive too frequent CE packets and respond to reach with CNP. These issues have already been observed with unicast and receivers limit the number of responses to CE packets already. Switches have the ability to control the probabilities of packets being marked with CE already.
2. Switches may receive too many CNPs. The switch can reduce the number of packets marked as CE to reduce CNPs to a desirable rate. The rate and logic of marking multicast packets as CE may differ from unicast packets and that may require switch vendors to introduce new tuneable parameters. Switches may use their knowledge of multicast replication trees to provide advanced means of congestion control.
3. Sender may receive too many CNPs. The sender may limit the number of CNPs that it reacts to in a given timeframe. It seems that there are already heuristics in place to avoid excessive throttling of senders for unicast.

Corporate Slides

Deutsche Börse Group

As an international exchange organisation and innovative market infrastructure provider, Deutsche Börse Group offers its customers a wide range of products, services and technologies covering the entire value chain of financial markets. It organises markets characterised by integrity, transparency and safety for investors who invest capital and for companies that raise capital.

Its business areas include the provision of index and ESG data, analytics and research solutions, trading and clearing services for investment instruments, securities settlement and custody, collateral and liquidity management, and investment fund services. In addition, the Group develops state-of-the-art IT solutions and offers IT systems all over the world.

With over 10,500 employees, the company has its headquarters in the financial centre of Frankfurt/Main, as well as a strong global presence in Luxembourg, Prague, London and Zug, in New York and Chicago, in Hong Kong, Singapore, Beijing, Tokyo and Sydney – and at other locations for its customers all over the world.

Contact

Corporate & Corporate Engagement

Deutsche Börse AG
Mergenthalerallee 61
65760 Eschborn

E-mail corporate.communications@deutsche-boerse.com

DEUTSCHE BÖRSE GROUP

Disclaimer

This publication is for informational purposes only. None of the information in this publication constitutes investment advice and does not constitute an offer to sell or a solicitation of an offer to purchase any contract, share or other financial instrument. This publication is not intended for solicitation purposes but only for use as general information. All descriptions, examples and calculations contained in this publication are for illustrative purposes only.

Deutsche Börse AG, Frankfurter Wertpapierbörse (FWB®, the Frankfurt Stock Exchange), Eurex Frankfurt AG, Eurex Deutschland and Eurex Clearing AG do not represent that the information in this publication is comprehensive, complete or accurate and exclude liability for any consequence resulting from acting upon the contents of this or another webpublication, in so far as no wilful violation of obligations took place or, as the case may be, no injury to life, health or body arises or claims resulting from the Product Liability Act are affected.

Securities traded on the Frankfurt Stock Exchange and Eurex derivatives (other than EURO STOXX 50® Index Futures contracts, EURO STOXX® Select Dividend 30 Index Futures contracts, STOXX® Europe 50 Index Futures contracts, STOXX® Europe 600 Index Futures contracts, STOXX® Europe Large/Mid/Small 200 Index Futures contracts, EURO STOXX® Banks Sector Futures contracts, STOXX® Europe 600 Banks/Industrial Goods & Services/Insurance/Media/Personal & Household Goods/Travel & Leisure/Utilities Futures contracts, Dow Jones Global Titans 50 IndexSM Futures contracts, DAX® Futures contracts, MDAX® Futures contracts, TecDAX® Futures contracts, SMIM® Futures contracts, SLI Swiss Leader Index® Futures contracts, RDXxt® USD - RDX Extended Index Futures contracts, Eurex inflation/commodity/weather/property and interest rate derivatives) are currently not available for offer, sale or trading in the United States nor may they be offered, sold or traded by persons to whom US tax laws apply.

The fund shares listed in XTF Exchange Traded Funds® are admitted for trading on the Frankfurt Stock Exchange. Users of this information service who legally reside outside Germany are herewith advised that sale of the fund shares listed in XTF Exchange Traded Funds may not be permitted in their country of residence. The user makes use of the information at their own risk.

Legal validity of this disclaimer

In the event that individual parts of or formulations contained in this text are not, or are no longer, legally valid (either in whole or in part), the content and validity of the remaining parts of the document are not affected.

Trademarks

The following names and designations are registered trademarks of Deutsche Börse AG or an affiliate of Deutsche Börse Group:

1585®; A7®; Buxl®; C7®; CDAX®; CEF®; CEF alpha®; CEF ultra®; CFF®; Classic All Share®; Clearstream®; CX®; D7®; DAX®; DAXglobal®; DAXplus®; DB1 Ventures®; DBIX Deutsche Börse India Index®, Deutsche Börse®; Deutsche Börse Capital Markets Partner®; Deutsche Börse Commodities®; Deutsche Börse Venture Network®; Deutsches Eigenkapitalforum®; DivDAX®; eb.rexx®; eb.rexX Jumbo Pfandbriefe®; ERS®; eTriParty®; Eurex®; Eurex Bonds®; Eurex Clearing Prisma®; Eurex Improve®; Eurex Repo®; Euro GC®; ExServes®; EXTF®; F7®; FDAX®; FWB®; GC Pooling®; GCPI®; GEX®; Global Emission Markets Access – GEMA®; HDAX®; iNAV®; L-DAX®; L-MDAX®; L-SDAX®; L-TecDAX®; M7®; MDAX®; N7®; ODAX®; ÖkoDAX®;PROPRIS®; REX®; RX REIT Index®; Scale®; SCHATZ-FUTURE®; SDAX®; ShortDAX®; StatistiX®; T7®; TecDAX®; Technology All Share®; TRICE®; USD GC Pooling®; VDAX®; VDAX-NEW®; Vestima®; Xcreen®, Xemac®; Xentric®, Xetra®; Xetra-Gold®; Xpect®; Xpider®; XTF®; XTF Exchange Traded Funds®; We make markets work®

The names and trademarks listed above do not represent a complete list and, as well as all other trademarks and protected rights mentioned in this publication, are subject unreservedly to the applicable trademark law in each case and are not permitted to be used without the express permission of the registered owner. The simple fact that this publication mentions them does not imply that trademarks are not protected by the rights of third parties.

The STOXX® indices, the data included therein and the trademarks used in the index names are the intellectual property of STOXX Ltd., Zug, Switzerland and/or its licensors. Eurex' derivatives based on the STOXX indices are in no way sponsored, endorsed, sold or promoted by STOXX and its licensors and neither STOXX nor its licensors shall have any liability with respect thereto.

STOXX iSTUDIO® is a registered trademark of STOXX Ltd., Zug, Switzerland.

TRADEGATE® is a registered trademark of Tradegate AG Wertpapierhandelsbank.

EEX® is a registered trademark of European Energy Exchange AG.

Flexible is better.® is a registered trademark of Axioma, Inc.