# ZTR-RTT Congestion Control Algorithm Overview

# Table of contents

# Overview

NVIDIA Zero Touch RoCE (ZTR) enables data centers to seamlessly deploy RDMA over Converged Ethernet (RoCE) without requiring any special switch configuration. Built according to the InfiniBand Trade Association (IBTA) RDMA standard and fully compliant with the RoCE specifications, ZTR enables seamless deployment of RoCE. ZTR also boasts performance equivalent to traditional switch-enabled RoCE and is significantly better than traditional TCP-based memory access. Moreover, with ZTR, RoCE network transport services operate side-by-side with non-RoCE communications in ordinary TCP/IP environments.

The new NVIDIA Congestion Control algorithm, Round-Trip Time Congestion Control (RTTCC) allows ZTR to scale to thousands of servers without compromising performance. Using ZTR and RTTCC allows data center operators to enjoy ease-of-deployment and operations together with the superb performance of Remote Direct Memory Access (RDMA) at a massive scale, without any switch configuration.

The new NVIDIA congestion control algorithm, RTTCC, actively monitors network RTT to proactively detect and adapt to the onset of congestion before dropping packets. RTTCC enables dynamic congestion control using a hardware-based feedback loop that provides dramatically superior performance compared to software-based congestion control algorithms. RTTCC also supports faster transmission rates and can deploy ZTR at a larger scale.

Main ZTR-RTT CC algorithm's characters are :

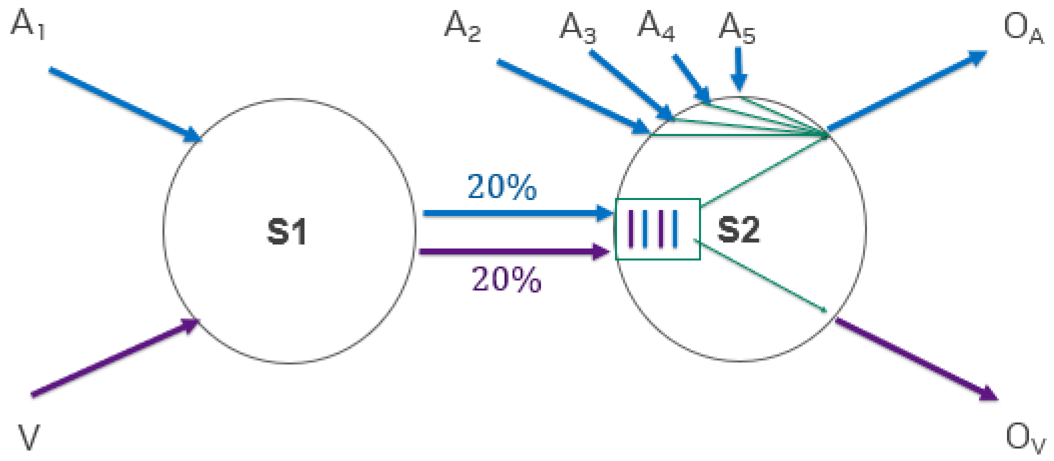
- Implemented on top of DPA (Data Path Accelerator)
- RTT-based congestion control
- Current default CC algorithm for RoCE
- Demonstrates better performance than DCQCN on HPC and AI workloads
- Maintain DCQCN good performance on storage workload

# Congestion Control Algorithm

Congestion Control provides performance isolation when multiple applications running on the same cluster. Additionally, it prevents congestion spreading when there is a slow receiver, reduce latency in the cluster, improves fairness, prevents parking-lot effects and packet's drop in lossy networks.

ZTR_RTTCC is NVIDIA’s default Congestion Control algorithm.

The diagram below shows an example of head of the line blocking scenario.



**Head of the Line Blocking Scenario**

## Datacenter Congestion Control Challenges

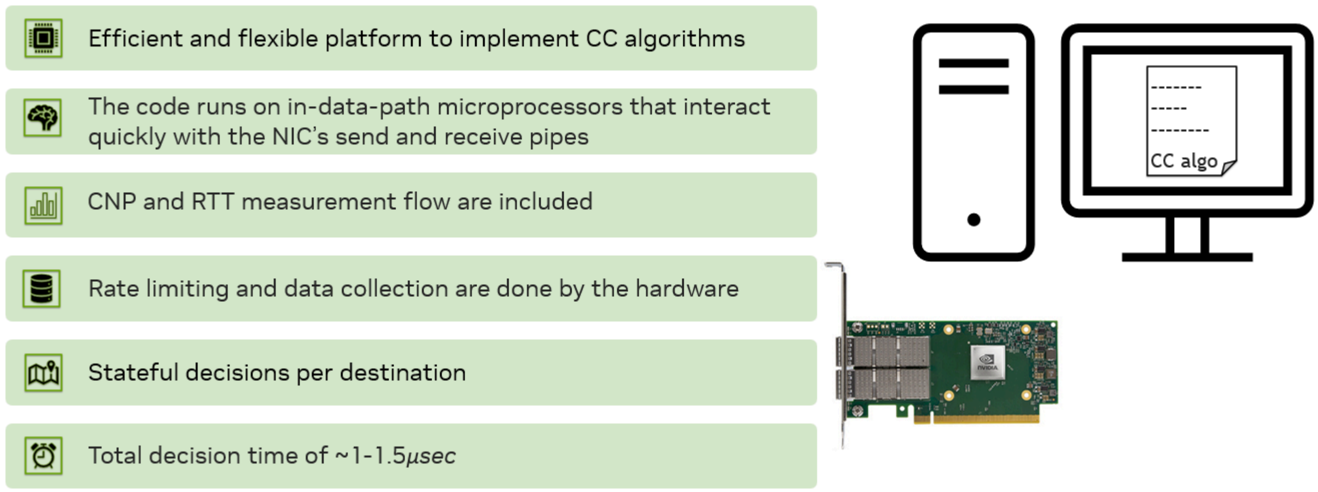
The following are the Datacenter Congestion Control challenges:

- Several µ-sec of latency with hundreds of Gbps of bandwidth
    - Congestion buildup is fast, so the congestion loop should be short
- A wide variety of traffic types, topologies and applications
    - Hard to develop an algorithm that suits all

- Congestion Control algorithms are constantly being introduced with new congestion indications

- Hardware implementation is not robust enough
- Software implementation reacts too slow

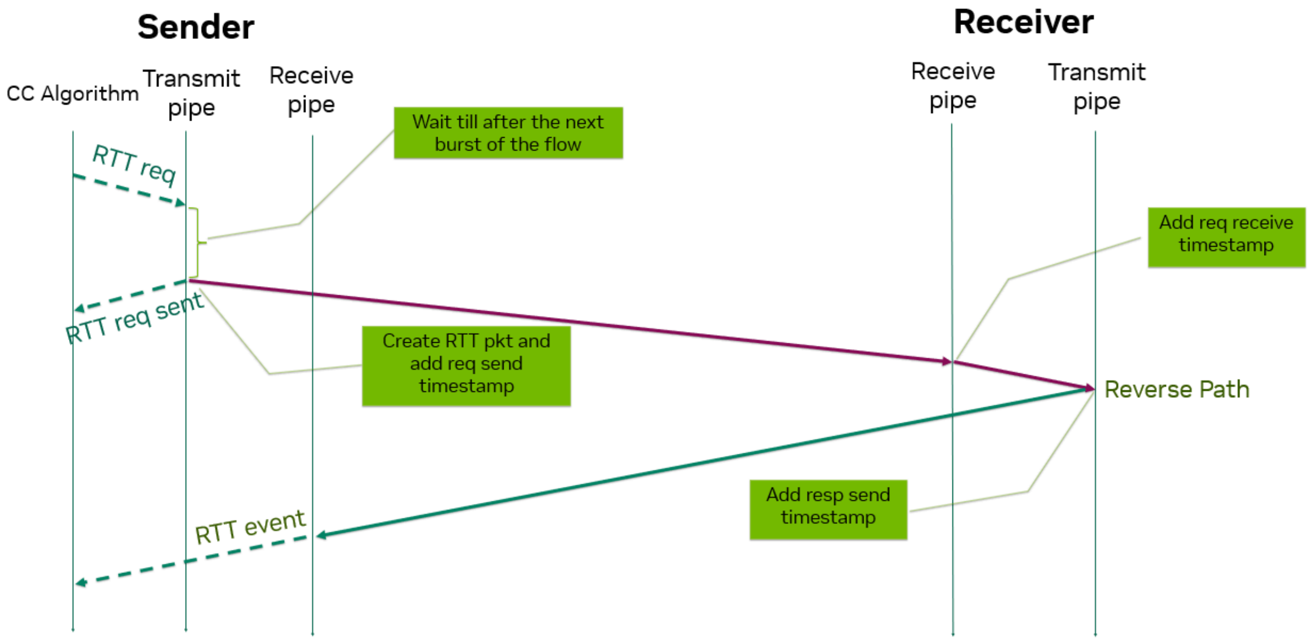
# ZTR RTTCC Infrastructure



- Efficient and flexible platform to implement CC algorithms
- The code runs on in-data-path microprocessors that interact quickly with the NIC's send and receive pipes
- CNP and RTT measurement flow are included
- Rate limiting and data collection are done by the hardware
- Stateful decisions per destination
- Total decision time of ~1-1.5$\mu sec$

### ZTR RTTCC Infrastructure

# RTT Measurement Flow



## RTT Measurement Flow

# ZTR RTTCC Algorithm

Congestion indications

- RTT
- CNP

Rate update scheme

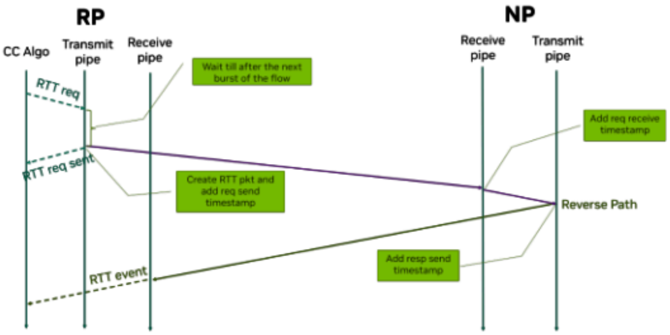
Evaluate congestion state by comparing RTT to $\frac{1}{\sqrt{rate}}$

Additive Increase

Multiplicative Decrease

Algorithmic windowing

Mimic window behavior by adjusting the rate when RTT changed



Fast reaction on first congestion

## ZTR RTTCC Algorithm

# Document Revision History

| Revision | Date | Description |
|---|---|---|
| 1.0 | September 25, 2024 | Initial release |

<b>Notice</b><br/><br/><br/><br/>This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.<br/><br/><br/><br/>NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.<br/><br/><br/><br/>Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.<br/><br/><br/><br/>NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.<br/><br/><br/><br/>NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.<br/><br/><br/><br/>NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.<br/><br/><br/><br/>No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.<br/><br/><br/><br/>Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.<br/><br/><br/><br/>THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.<br/><br/><br/><br/><b>Trademarks</b><br/><br/><br/><br/>NVIDIA and the NVIDIA logo are trademarks and/or registered trademarks of