IP Performance Measurement
Internet-Draft
Intended status: Standards Track
Expires: 22 December 2025

G. Fioccola
T. Zhou
Huawei
G. Zhao
Z. Li
China Mobile

20 June 2025

Data Fields for Congestion Measurement
draft-shi-ippm-congestion-measurement-data-04

Abstract

Congestion Measurement collects the congestion information in the packet while the packet traverses a path. The sender sets the congestion measurement data fields in the packet header indicating the network device along the path to update the congestion information field in the packet. When the packet arrives at the receiver, the congestion information field will reflect the degree of congestion across network path. Congestion Measurement can enable precise congestion control, assist in effective load balancing, and simplify network debugging. This document defines data fields for Congestion Measurement. Congestion Measurement Data-Fields can be encapsulated into a variety of protocols, such as IPv6, Segment Routing Header (SRH), Network Service Header (NSH), Generic Network Virtualization Encapsulation (Geneve), etc.

Status of This Memo

Table of Contents

1. Introduction

To effectively manage network congestion, a detailed understanding of congestion levels across the network is needed. Congestion control algorithms, therefore, necessitate precise congestion measurements to adapt and optimize data flow. This approach involves monitoring various metrics such as packet loss, delay variations, and throughput, which can provide an idea of the network's congestion state. Enhanced congestion metrics allow for a suited response to congestion, enabling algorithms to adjust sending rates with greater precision, thereby improving overall network performance and efficiency.

Furthermore, the detailed congestion measurements obtained are not solely beneficial for congestion control; they serve multiple purposes, including load balancing and network operations debugging. By analyzing congestion data, network operators can identify and

resolve bottlenecks, optimize traffic distribution, and ensure a balanced load across the network. This data-driven approach facilitates proactive network management, allowing for timely interventions that can preempt potential disruptions and enhance network reliability and performance.

High Precision Congestion Control (HPCC)[I-D.draft-miao-ccwg-hpcc], leverages INT (Inband Network Telemetry) for detailed congestion signal collection but faces challenges with packet size increases and computational redundancy. This document proposed a different approach and introduces data fields for Congestion Measurement. Congestion Measurement expands the conventional single-bit ECN to multiple bits, allowing network devices to update congestion information at each hop more granularly. Consequently, when packets reach the receiver, the congestion information field in the packet indicates not only the presence of congestion but the degree of congestion across the link's path. This approach facilitates a richer set of data for decision-making, supporting not only more precise congestion control but also improving load balancing and network debugging efforts. By overcoming HPCC's shortcomings, our approach enhances network efficiency, reduces computational overhead at endpoints, and offers a scalable solution to managing congestion in complex network environments. Congestion Measurement Data-Fields can be encapsulated into a variety of protocols, such as IPv6, Segment Routing Header (SRH), Network Service Header (NSH), Generic Network Virtualization Encapsulation (Geneve), etc.

1.1. Terminology

* CC: Congestion Control
* DRE: Discounting Rate Estimator[CONGA]
* ECN: Explicit Congestion Notification
* HPCC: High Precision Congestion Control[I-D.draft-miao-ccwg-hpcc]

1.2. Requirements Language

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.

2. Overview

Figure 1 shows the overview procedure of Congestion Measurement. First the sender MUST sets the packet with data fields for Congestion Measurement (see Section 3) which specifies what kind of the congestion information that the sending node intends to collect from transit nodes. As the packet traverses through the network, each router should inspect the data fields and update the Congestion Info Data field accordingly. Upon reaching the receiver, the updated congestion info data within the packet is extracted and then send back to the sender. The sender, now equipped with the congestion information reflective of the packet's journey, uses this data to make proper adjustments to its sending rate or load balancing decisions.

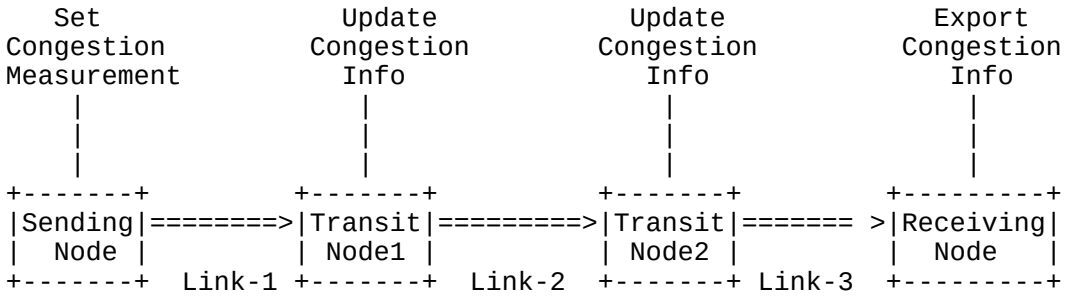


Figure 1: Overview of Congestion Measurement

3. Data fields for Congestion Measurement

Figure 2 shown the format of data fields for Congestion Measurement.

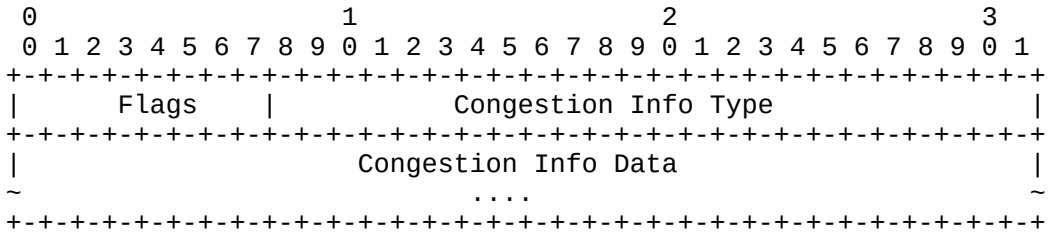
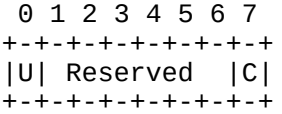


Figure 2: Data Fields for Congestion Measurement

The Flags field is shown below:



where:

* Flags: An 8-bit field.

  - The first bit(U) indicates whether the Congestion Info Data field needs to be updated by transit nodes. If set, the transit nodes will update the Congestion Info Data. If not, the transit node will not update it.

  - The last bit(C) indicates that the Congestion Info Data is customized and used only in limited domain such as Data Center network. If the C is 0, the Congestion Info Type is a bitmap. Other bits are reserved.

* Congestion Info Type: A 24-bit map that specifies the present Congestion Info Data. Supported Congestion Info Data is listed in Table 1. Note that it is possible for multiple Congestion Info Data to coexist in one packet for the endpoint to collect the detailed raw congestion information.

* Congestion Info Data: A variable length field including the congestion information data. Router MUST update this field based on local load status. The length and the update operation is listed in Table 1.

The Congestion Info Type is a 24-bit identifier which specifies which data types are used in this data list. It is a bit field and the following bits are defined in this document. The order of packing the data fields in the Congestion Info Data follows the bit order of the Congestion Info Type field, as follows:

Bit	Congestion Info Data	Length	Operation
0	Inflight Ratio	8	Max
1	DRE	8	Max
2	Queue Utilization Ratio	8	Max
3	Queue Delay	8	Add
4	Congested Hops	8	Add
5	Available Bandwidth	8	Min

Table 1: Congestion Info Data

The Congestion Info Data fields are:

* Inflight Ratio is described in [I-D.draft-miao-ccwg-hpcc] and further explained in Section 4.

* DRE is defined in [CONGA]. It is a simple module for measuring the load of a link. The DRE maintains a register, which is incremented for each packet sent over the link by the packet size in bytes, and is decremented periodically with a multiplicative factor between 0 and 1. Therefore, the register is proportional to the rate of traffic over the link. The congestion metric for the link is obtained by comparing it with the link speed.

* Queue Utilization Ratio is calculated as the arrival rate divided by the service rate.

* Queue Delay is the time a packet waits in a queue before being processed or transmitted.

* Congested Hops is the number of congested hops along the path.

* Available Bandwidth is further explained in Section 5.

4. HPCC with Inflight Ratio

As described in [I-D.draft-miao-ccwg-hpcc], HPCC calculates the inflight ratio of each link(represent the link utilization of the link) from the collected raw load information carried in the p4.org INT. Then maximum inflight ratio along the path is identified and used to adjust the sending rate. The formula to calculate the inflight ratio of each link is shown below:

txRate = (txBytes_1 - txBytes_2)/(t_1-t_2)
inflight ratio = qlen/(B*T) + txRate/B

where:

* txBytes: link total transmitted bytes associated with timestamp ts

* qlen: link queue length

* B: link bandwidth

* T: Baseline RTT

Leveraging Congestion Measurement, the router participates in calculation of the maximum inflight ratio. Each router MUST calculate the inflight ratio of the down link and then compare it to the one in the Congestion Info Data field and keep the larger one. When the packet arrives at the endpoint, the Congestion Info Data field already contains the maximum inflight ratio. The sending rate adjustment algorithm remains unchanged. By allowing routers to conduct these calculations, the computing overhead is reduced for the endpoint. Since the value is updated at each hop, the packet size remains unchanged independently from the number of hops.

5. Congestion Control with Available Bandwidth

The ABW (Available Bandwidth) of links can be applied in existing CC algorithms to optimize their throughput performance, such as TCP Reno and CUBIC. The sending rate and congestion window can be dynamically adjusted during the CC’s slow-start and loss recovery phases. The BBR algorithm, which detects link bottleneck bandwidth based on rate and round-trip time (RTT), can utilize the ABW to obtain the bottleneck bandwidth of the link and optimize data throughput efficiency. Alternatively, a completely new CC algorithm can be designed based on ABW to predict and avoid congestion in advance.

The method for obtaining the ABW of a link is shown as follows:

1. The sending node can obtain the ABW of its egress port, mark the packet with data fields for ABW Measurement, and then send the packet to the Receiving node.

2. The transit node identify the ABW probe action based on the Congestion Measurement header, compare the ABW of their egress port with the ABW in the packet. If the ABW of the current node is smaller than that in the packet, it updates to the link's ABW and forwards the packet; otherwise, it directly forwards the packet.

3. After receiving the ABW packet, the receiving node parses the link's ABW, constructs an ABW response packet, and sends it back to the sending node.

The calculation of the current node's ABW can be referenced as follows:

ABW = B - T - R

where:

* B is the bandwidth of the egress port where the flow passes

* T is the traffic size of that egress port

* R is the reserved bandwidth

The reserved bandwidth takes into account the fairness of the CC algorithm, facilitating the entry of newly added flow. The value of R can be set according to the specific circumstances of each node, allowing TOR switches and backbone routers to reserve different percentages of bandwidth.

6. Security Considerations

An attack on Congestion Measurement can prevent the collection of the congestion information by maliciously modifying the data fields in transit or by injecting packets with maliciously generated data fields. As mentioned above, a possibility to overcome this issue can be to apply the Congestion Measurement in specific controlled domains, thus confining the potential attack vectors within the network domain. A limited administrative domain provides the network administrator with the means to select, monitor, and control the access to the network, making it a trusted domain.

7. IANA Considerations

IANA is requested to define a registry group named "Congestion Measurement Parameters". This registry defines code points for Flags, Congestion Info Type and Congestion Info Data as explained in Section 3.

8. References

8.1. Normative References

[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/rfc/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/rfc/rfc8174>.

8.2. Informative References

[CONGA] Alizadeh, M., Edsall, T., Dharmapurikar, S., Vaidyanathan, R., Chu, K., Fingerhut, A., Lam, V., Matus, F., Pan, R., Yadav, N., and G. Varghese, "CONGA: distributed congestion-aware load balancing for datacenters", ACM, Proceedings of the 2014 ACM conference on SIGCOMM pp. 503-514, DOI 10.1145/2619239.2626316, August 2014, <https://doi.org/10.1145/2619239.2626316>.

[I-D.draft-miao-ccwg-hpcc] Miao, R., Anubolu, S., Pan, R., Lee, J., Gafni, B., Tantsura, J., Alemania, A., and Y. Shpigelman, "HPCC++: Enhanced High Precision Congestion Control", Work in Progress, Internet-Draft, draft-miao-ccwg-hpcc-03, 6 January 2025, <https://datatracker.ietf.org/doc/html/draft-miao-ccwg-hpcc-03>.

Contributors

Hang Shi
Huawei
Beijing
China
Email: shihang9@huawei.com

Authors' Addresses

Giuseppe Fioccola
Huawei
Vimodrone (Milan)
Italy
Email: giuseppe.fioccola@huawei.com

Tianran Zhou
Huawei
Beijing
China
Email: zhoutianran@huawei.com

Guangyu Zhao
China Mobile
China
Email: zhaoguangyu@chinamobile.com

Zhenqiang Li
China Mobile
Beijing
China
Email: li_zhenqiang@hotmail.com