MPTCP Working Group
Internet-Draft
Intended Status: Standards Track
Expires: April 29, 2018

F. Wang
J. Zuo
Z. Cao
K. Zheng
Huawei

October 30, 2017

A Proactive Approach to Avoid Performance Degradation of MPTCP
draft-zuo-mptcp-degradation-00

Abstract

One of the goals for MPTCP is utilizing multiple paths to perform at least as well as the best path in terms of throughput. However, this goal might not be arrived at because of the path asymmetry, which is called as the performance-degradation problem of MPTCP in this draft. Some existing methods focus on this problem, such as penalizing and opportunistic retransmission, which reactively responds to the head-of-line blocking for trying their best to send data across all paths. In order to efficiently utilize the capabilities of the multiple paths, this draft proposes an approach that proactively selects the best path(s) to send data instead of always bonding all paths together.

Table of Contents

1. Introduction

MultiPath TCP (MPTCP) enables a transport connection across multiple paths simultaneously [RFC6824]. According to [RFC6356], one of the MPTCP's goals is improving throughput: a multipath flow should perform at least as well as a single path flow would on the best of the path available to it. However, this goal cannot be always achieved due to the head-of-line blocking caused by the path asymmetry, e.g. WiFi and LTE in smart phones. To be convenient, this phenomenon (that MPTCP performs worse than the best path) is called as the performance-degradation problem in this draft.

The direct solution for this problem is allocating a large enough receive buffer. As in [RFC6182], 'The RECOMMENDED receive buffer is

2*sum(BW_i)*RTT_max, which ‘ensures subflows do not stall when fast retransmit is triggered on any subflow’. However, the buffer size can be very large to cover all the possible scenarios. In other words, the buffer size can be limited and the performance-degradation problem is possible to exist. Therefore, it needs a solution for MPTCP protocol that can solve this problem. On this issue, some reactive methods have been proposed, such as penalizing and opportunistic retransmission. They take actions after the head-of-line blocking occurring and their purpose is to send data across all paths as possible as they can. Experiments show that even with these methods, the capabilities of the multiple paths may not be efficiently utilized. Meanwhile, instead of bonding all paths together, [RFC6182] indicates that it would be better ‘to only use some of the fastest available paths for the MPTCP connection in extreme cases’.

This draft focuses on this problem and proposes a proactive approach, which dynamically employs part or all of the paths for higher utilization efficiency. In particular, this approach first measures the aggregated throughput of the multiple paths and the throughput of the best single path in real time. Then, if the performance-degradation phenomenon is derived through the throughput comparison, only the best path is used to send data.

2. Acronyms and Terminology

MPTCP: Multiple Path Transport Control Protocol

RTT: Round-robin Transmit Time

PLR: Packet Loss Ratio

BW: BandWidth

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].

3. A Proactive Approach to Avoid MPTCP Performance Degradation

This section introduces the basic principles of the proposed approach: 1) how to efficiently measure the characteristics of multiple paths; 2) how to select the path(s). More details regarding to this approach can be found in Section 4. Briefly, through measuring and comparing the path characteristics, the best path(s) is selected to satisfy a special purpose, such as the high throughput or

low RTT. In theory, path characteristics can be RTT, throughput, congestion window, etc., while only throughput is considered in this draft for a case study. As shown in Figure 1, the proposed approach is implemented in the Performance-Degradation-Free Module of the MP layer.

3.1 Throughput Measurement

This section describes the method of the throughput measurement for multipath transmissions. As shown in [CMT-SCTP], one possible method is modeling the aggregated throughput of multiple paths and measuring the related path characteristics, i.e. buffer size, PLR, BW and RTT. However, it is hard to accurately model the throughput, since 1) lacking of BW evaluation methods, 2) and the small PLR can be obtained only after a huge number of data exchanging.

This draft proposes a method that directly measures the multipath and best-single-path throughputs, in order to avoid the measurements of BW and PLR. At first, two modes are defined for the throughput measurement:

1) Redundant transmission mode: Multiple paths transmit the same data;

2) MultiPath (MP) transmission mode: Multiple paths transmit different data with using a MPTCP scheduling scheme, such as minRTT (the default scheduling method) in version 0.92. By the way, the opportunistic retransmission and penalization mechanisms can be enabled in MP mode.

The redundant transmission mode is a kind of scheduling scheme in MPTCP v0.92, which is usually used to achieve lower packet delay than that of a single path. Because of wasting bandwidth (BW), the redundant transmission mode is ignored when MPTCP aggregates the BWs of multiple paths. However, this draft utilizes the redundant transmission mode to measure the throughputs of each path and the best single path, while using the multipath transmission mode to measure the aggregated throughput. To be convenient, the redundant/MP transmission mode is also called as the redundant/MP mode in the following part of this draft. Moreover, the throughput measurement is executed during the data transmission without introducing extra packets, and is periodic for self-adapting to the dynamic network environment.

The Figure 1 briefly shows the relationship between the two transmission modes and our Performance-Degradation-Free Module in MP layer.

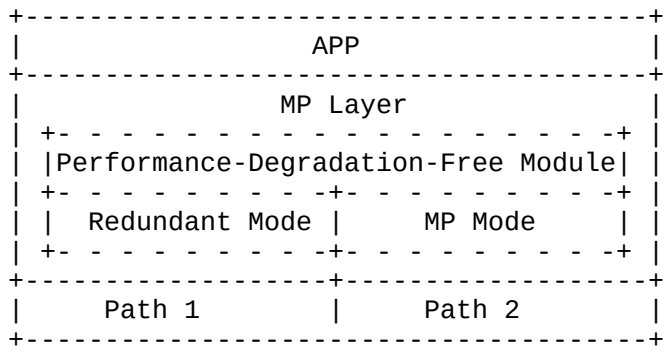

Figure 1: The performance-degradation-free module in the MP layer.

3.2 Path Selection Strategy

All these paths or the best one of them are used to send data depending on the corresponding throughput.

4. Operation Overview

Section 3 presents the main principles, i.e. throughput measurement and path selection. This section introduces how the proposed approach works in details. According to different congestion-control schemes, the proposed approach may have slow-start stage and congestion-control stage. At the first stage, the approach ensures that the slow start of the scheme has been finished. At the second stage, the approach periodically measures the average throughputs of multiple paths and the best path, and employs the suitable path(s) through throughput comparisons.

4.1 Slow-Start Stage

During the slow-start stage, MPTCP schedules the data in the redundant mode, where the same data are transmitted across multiple paths. After each round, we get a measured throughput from the view of MP layer, where the round time could be from sending a packet until receiving its responsive ACK. The throughput (defined as ‘B_rd’) is calculated by dividing the total number of delivered data with the time period of this round. We define a threshold ‘h’, where h belongs to (0,1), e.g. h=0.2. As shown in Figure 2, if the measured throughput is h times larger than the last measured value (i.e. B_rd(i) >= h * B_rd(i-1)) , keep the redundant mode and set the round counter Nr as 0, or else increase the round counter (defined as Nr) by 1.

4.2 Congestion-Avoidance Stage

If the round counter Nr > N (e.g. N = 3), then the transmission switches to the congestion-avoidance stage, as shown in Figure 2. Two timers are set to decide when we need measure the throughputs of the redundant and MP modes again. Two indicators of updating the redundant and MP measurement values, Indicator_Redundant and Indicator_MP, are both set as 1. The indicator timers for the redundant mode and MP modes are set as REDUNDANT_MODE_TIMER and MP_MODE_TIMER, separately. The timeout of both timers is set as a period of time TIMEOUT (e.g. 10s).

At first, the redundant mode is used to measure the throughput.

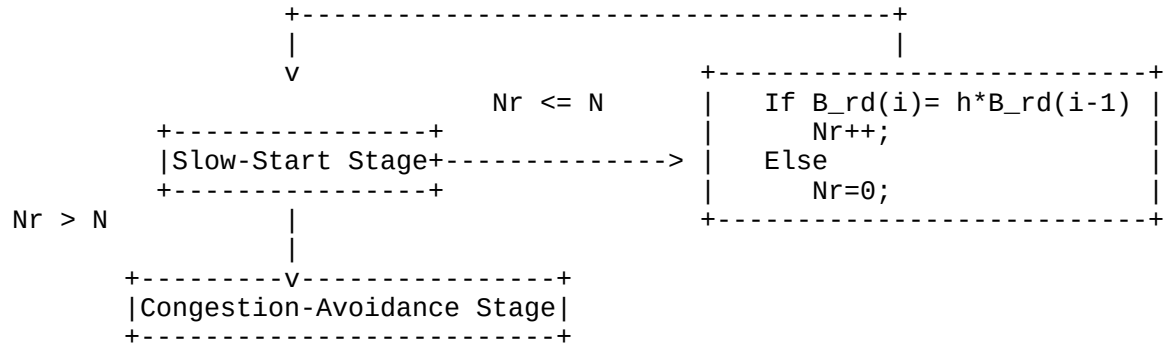

Figure 2: Stage transition from slow start to congestion avoidance.

4.2.1 Redundant Transmission Mode

As shown in Figure 3, when running the redundant mode, MP layer and each path would calculate the average throughputs during a period of time (e.g. 1s). Meanwhile, the timer REDUNDANT_MODE_TIMER is reset and the variable Indicator_Redundant is set as 0. Before sending each segment, if Indicator_MP = 1, the redundant mode is switched to the MP mode to measure the throughput, as described in Section 4.2.2. Otherwise, the two throughputs obtained from the redundant and MP modes are compared, which is introduced in Section 4.2.3.

Figure 3: The redundant-and-MP mode transition when the corresponding indicator is equal to 1.

4.2.2 Multipath Transmission Mode

As shown in Figure 3, when running the MP mode, MP layer would calculate its average throughput during a period of time (e.g. 1s). Meanwhile, the timer MP_MODE_TIMER is reset and the variable Indicator_MP is set as 0. Before sending each segment, if Indicator_Redundant = 1, the MP mode is switched to the redundant mode to measure the throughputs (refer to Section 4.2.1). Otherwise,

the two throughputs obtained from the redundant and MP modes are compared (see Section 4.2.3).

4.2.3 Throughput Comparison

By comparing the throughputs measured by the redundant and MP modes, a transmission mode corresponding to the larger throughput would be selected.

If the redundant mode (described in Section 4.2.1) is selected, the

path with the highest throughput is employed for data transmission. Before sending each segment, if Indicator_redundant = 1 or Indicator_MP = 1, we will go back to the redundant mode or the MP mode to measure the corresponding throughput(s). If both of the indicators are 0, keep transmitting data at the best path.

If the MP mode (described in Section 4.2.2) is selected, the multiple paths are bonded together to achieve the aggregated throughput.

4.2.4 Indicator Timer Timeout

During data transmission, if the timer for the redundant/MP mode is timeout, set Indicator_Redundant/Indicator_MP = 1 and reset the timer of the redundant/MP mode.

5. Security Considerations

TBD.

6. IANA Considerations

7. References

7.1. 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, <http://www.rfc-editor.org/info/rfc2119>.

[RFC6356] Raiciu, C., Handly, M., and Wischikf, D., "Coupled Congestion Control for Multipath Transport Protocols", RFC 6356, DOI 10.17487/RFC6356, October 2011, <https://rfc-editor.org/rfc/rfc6356.txt>.

[RFC6824] Ford, A., Raiciu, C., Handley, M., and Bonaventure, O., "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, DOI 10.17487/RFC6824, January 2013, <http://www.rfc-editor.org/info/rfc6824>.

[RFC6182] Ford, A., Raiciu, C., Handley, M., Barre, S., Iyengar, J., "Architectural Guidelines for Multipath TCP Development", RFC 6182, DOI 10.17487/RFC6182, March 2011,

<http://www.rfc-editor.org/info/rfc6182>.

7.2. Informative References

[CMT-SCTP] Yang, W., Li, H., Li, F., Wu, Q., & Wu, J., "RPS: range-based path selection method for concurrent multipath transfer", June 2010, In Proceedings of the 6th International Wireless Communications and Mobile Computing Conference (pp. 944-948). ACM.

Author’s Addresses

Fanzhao Wang
Huawei Technologies
Bantian, Longgang District,
Shenzhen 518129 P.R. China
EMail: wangfanzhao@huawei.com

Jing Zuo
Huawei Technologies
Bantian, Longgang District,
Shenzhen 518129 P.R. China
EMail: jing.zuo@huawei.com

Zhen Cao
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China
EMail: zhencao.ietf@gmail.com

Kai Zheng
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China
EMail: kai.zheng@huawei.com