Abstract

This document describes the mechanism of wVegas (weighted Vegas), which is a delay-based congestion control for MPTCP. The current congestion control algorithm of MPTCP, LIA, achieves only course-grained load balancing, since it is based on packet loss event. On the contrary, wVegas adopts packet queuing delay as congestion signals, thus achieving fine-grained load balancing. Compared with loss-based algorithms, wVegas is more sensitive to changes of network congestion and thus achieves more timely traffic shifting and quicker convergence. WVegas has been implemented in the Linux Kernel and will be a part of UCLouvain’s MPTCP implementation.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on January 4, 2015.
1. Introduction ................................................ 3
   1.1. Requirements Language................................ 4
   1.2. Requirements Language................................ 4
2. Weighted Vegas Algorithm.................................... 6
3. Practical considerations.................................... 8
4. Discussion ................................................ 9
5. Security Considerations.................................... 10
6. IANA Considerations ...................................... 10
7. References ................................................ 10
   7.1. Normative References................................. 10
   7.2. Informative References............................... 10
1. Introduction

Performing congestion control independently on each path cannot guarantee the fairness for multipath transportation. So the major goal of multipath congestion control is to couple all the subflows belonging to a flow in order to achieve both fairness and efficiency. The current MPTCP adopts a congestion control algorithm called Linked Increases algorithm [RFC6356]. It provides the ability to shift traffic from more congested paths to less ones. However, it achieves only course-grained load balancing, because it uses the packet losses as the signal of network congestion and will shift traffic only after the loss event. Other alternative congestion control algorithms, such as OLIA [OLIA], have the same way to judge the congestion. These proposals lack the finer-grained information related to the degree of congestion.

In this draft, we introduce weighted Vegas (wVegas), which is a delay-based multipath congestion control algorithm. Comparing to LIA and OLIA, wVegas adopts packet queuing delay as congestion signals, thus achieving fine-grained load balancing. [WVEGAS] WVegas is developed using the network utility maximization model [ADMTQM]. By solving the maximization problem, we get a general framework for designing an algorithm of multipath congestion control, which constitutes the Congestion Equality Principle and an approximate iterative algorithm [WVEGAS]. The Congestion Equality Principle illustrates that a fair and efficient traffic shifting implies every flow strives to equalize the extent of congestion that it perceives on all its available paths. And the approximate iterative algorithm makes the solution practical in real networks. The wVegas is precisely derived from the framework.

As the name shows, wVegas is originated from TCP-vegas [VEGAS] that measures packet queuing delay to estimate the extent of network congestion. TCP-vegas has two configurable parameters alpha and beta, for adjusting the congestion window during the congestion avoidance phase. Since the two parameters are commonly very close to each other, wVegas uses only one parameter for brevity. The design philosophy of wVegas is depicted as follows. First, wVegas performs in the same way as TCP-vegas on each path. Second, each flow has a fixed sum of parameters alpha, in spite of the number of subflows the flow has. Third, wVegas adaptively adjusts the parameter alpha resulting in the change of the transmission rate of the related subflows so as to equalize the extent of congestion on the path.

WVegas has been implemented in the Linux Kernel and will be a part of UCLouvain’s MPTCP implementation [MPLKI]. In [VEGAS], we study the
traffic shifting ability of wVegas, reveal that it can guarantee the intra-protocol fairness and show the domino effect which is an indication of good performance.

1.1. Requirements Language

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 [RFC2119].

1.2. Requirements Language

Regular TCP: The standard version of TCP, which is currently restricted to a single path per connection.

Multipath TCP (MPTCP): A modified version of the regular TCP that provides the ability to simultaneously use multiple paths between peers.

LIA: The Linked-Increases Algorithm of MPTCP. [RFC6356]

OLIA: The Opportunistic Linked-Increases Algorithm for MPTCP. [OLIA]

WVegas: Weighted Vegas, which is a delay-based multipath congestion control algorithm for MPTCP. [WVEGAS]

Subflow (path): A single path TCP connection always belonging to a MPTCP connection.

Expected sending rate: The best rate that a single path flow can get under the current congestion window condition.

Actual sending rate: The actual rate that a single path flow can get under the current congestion window condition.

Diff_r: For subflow r, the difference between Expected sending rate and Actual sending rate.

Cwnd_r: The congestion window on a subflow r (maintained in packets).

Ssthresh_r: The slow start threshold of a subflow r.

Rtt_r: The average RTT in the last round on a subflow r.

Rate_r: The rate of subflow r, which is used to estimate the equilibrium rate of subflow r.
Base_rtt_r: The RTT of a subflow r when the sub-connection is not congested.

Alpha_r: TCP-vegas tries to keep the extra bytes between two configurable parameters in a single path flow. For subflow r, we use only one parameter, alpha, for brevity.

Total_alpha: The total bytes backlogged in the network for all subflows belonging to one MPTCP flow.

Weight_r: The rate of subflow r divided by the total rate among all subflows belonging to one MPTCP flow.

Gamma: When the Actual sending rate falls below the Expected sending rate by the gamma threshold, TCP-vegas changes from slow start to congestion avoidance phase.

Queue_delay_r: For subflow r, queue_delay_r records the minimal queuing delay measured after the last backoff.
2. Weighted Vegas Algorithm

In this section, we introduce wVegas. wVegas is a delay-based congestion control algorithm and also uses the unmodified TCP behavior in the case of loss event. wVegas is an alternative for LIA, the current congestion control of MPTCP.

The algorithm mainly applies to the congestion avoidance phase and, in slow start phase, it also add a chance to enter the congestion avoidance phase earlier, which is similar with the implementation of the TCP-vegas [VEGAS]. The decrease of the congestion avoidance phase, the fast retransmit and fast recovery algorithms are the same as TCP [RFC5681].

The following operations of wVegas must be performed on the end of each transmission round.

For a subflow \( r \), the difference between the Expected sending rate and Actual sending rate is calculated as follows:

\[
\text{diff}_r = (\text{cwnd}_r / \text{base}_rtt_r - \text{cwnd}_r / \text{rtt}_r) \times \text{base}_rtt_r
\]

If the subflow is in the slow start phase and the \( \text{diff}_r \) is larger than \( \text{gamma} \), it must enter the congestion avoidance phase. This operation is inherited from TCP-vegas [VEGAS] and can be achieved by setting the \( \text{ssthresh}_r \) to \( (\text{cwnd}_r - 1) \). On the other hand, if the \( \text{diff}_r \) is no more than \( \text{gamma} \) in slow start phase, wVegas must act the same way as TCP [RFC5681].

In the congestion avoidance phase, if the \( \text{diff}_r \) is no less than \( \text{alpha}_r \), the \( \text{rate}_r \) must be updated. The \( \text{weight}_r \) and the \( \text{alpha}_r \) must be tweaked before the window adjustment and the improvement of the \( \text{base}_rtt_r \) accuracy. The \( \text{rate}_r \), the \( \text{weight}_r \) and the \( \text{alpha}_r \) must be calculated as follows:

\[
\text{rate}_r = \frac{\text{cwnd}_r}{\text{rtt}_r} \quad (1)
\]

\[
\text{weight}_r = \frac{\text{rate}_r}{\sum(\text{rate}_r)} \quad (2)
\]

\[
\text{alpha}_r = \text{weight}_r \times \text{total}_\alpha
\]

\( \sum(\text{rate}_r) \) is the total rate of all subflows belonging to one MPTCP flow. After the tweak of the \( \text{alpha}_r \), if the tweak is needed, the \( \text{cwnd}_r \) must be adjusted as follows:

- If \( \text{diff}_r \) is larger than \( \text{alpha}_r \), then
cwnd_r = cwnd_r - 1

- If diff_r is less than alpha_r, then
  
cwnd_r = cwnd_r + 1

The last task wVegas has to do is to try to drain link queues. This operation is to improve the accuracy of base_rtt_r. And the specific method is described in [DRLK]. The idea is to make congestion window back off once detecting the queuing delay is larger than some threshold, so that the bottleneck link can drain off the backlogged packets. And thus all the flows involved have a chance to obtain the more accurate propagation delay.

First, wVegas has to calculate the current queuing delay as follows:

\[
\text{queue\_delay}_r = \text{rtt}_r - \text{base\_rtt}_r
\]

If the current queuing delay is less than the saved queue\_delay\_r, the queue\_delay\_r must be replaced by the current one. And if the current queuing delay is two times larger than queue\_delay\_r, the following operations must be performed:

\[
cwnd_r = cwnd_r * 0.5 * \frac{\text{base\_rtt}_r}{\text{rtt}} \quad (3)
\]
3. Practical considerations

In practice, it is difficult to get the RTT of a subflow $r$ when the sub-connection is not congested. WVegas should set the $\text{base}_\text{rtt}_r$ to the minimum of all the measured round trip times. The $\text{total}_\text{alpha}$ should be implemented as a configurable parameter.

Equation (1) and (2) imply that the rate and the weight are floating point values. However, in many kernels, floating point operations are disabled. There is an easy way to approximate the above calculations using integer arithmetic.

Let $\text{rate}_\text{scale}$ be an integer. When computing the rate, use $\text{cwnd}_r \times \text{rate}_\text{scale}$ instead of $\text{cwnd}_r$ and the related operations can be done in integer arithmetic. The $\text{rate}_r$ should be calculated as follows:

$$\text{rate}_r = \text{cwnd}_r \times \text{rate}_\text{scale} / \text{rtt}_r$$

The $\text{rate}_\text{scale}$ implies the precision we need for computing the rate. With this change, the rate can be stored as an integer variable. Besides, when we need to calculate the sum rate of all subflows belonging to one flow, the operations can be done in integer arithmetic.

When updating the $\text{weight}_r$, we also need a $\text{weight}_\text{scale}$ to avoid floating point operations. So the $\text{weight}_r$ should be computed as follows:

$$\text{weight}_r = \text{rate}_r \times \text{weight}_\text{scale} / \text{SUM}(\text{rate}_r)$$

The $\text{weight}_\text{scale}$ supplies a similar function with $\text{rate}_\text{scale}$. With the $\text{weight}_\text{scale}$, $\text{alpha}_r$ can be much more accurate. But it also need scale down as follows:

$$\text{alpha}_r = \text{weight}_r \times \text{total}_\text{alpha} / \text{weight}_\text{scale}$$

For the sake of brevity, we combine the $\text{rate}_\text{scale}$ and $\text{weight}_\text{scale}$ to one scale parameter. We name it $\text{wvegas}_\text{scale}$. It would be better to set the $\text{wvegas}_\text{scale}$ as a power of two, which allows faster shift operations rather than multiplication and division.

In equation (3), the multiplication by 0.5 can be implemented by shift operation.
4. Discussion

Congestion Equality Principle shows that a fair and efficient traffic shifting implies that every flow strives to equalize the extent of congestion that it perceives on all its available paths. This principle has been proved in [WVEGAS]. By instantiating the approximate iterative algorithm, weighted Vegas (wVegas), a delay-based algorithm for multipath congestion control, was developed, which uses packet queuing delay as congestion signals, thus achieving fine-grained load balancing. Our simulations show that, compared with loss-based algorithms, wVegas is more sensitive to changes of network congestion and thus achieves more timely traffic shifting and quicker convergence. Additionally, as it occupies fewer link buffers, wVegas rarely causes packet losses and shows better intra-protocol fairness.
5. Security Considerations

Security considerations discussed in [RFC6181] and [RFC6356] are to be taken into account.

6. IANA Considerations

This document has no actions for IANA.

7. References

7.1. Normative References


7.2. Informative References


Authors’ Addresses

Mingwei Xu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China
Phone: +86-10-6278-5822
EMail: xmw@cernet.edu.cn

Yu Cao
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China
Phone: +86-185-3927-5620
Email: caoyu08@csnet1.cs.tsinghua.edu.cn

Enhuan Dong
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China
Phone: +86-183-1105-7156
Email: deh13@mails.tsinghua.edu.cn