NAME

cc_dctcp - DCTCP Congestion Control Algorithm

DESCRIPTION

The DCTCP (data center TCP) congestion control algorithm aims to maximise throughput and minimise latency in data center networks by utilising the proportion of Explicit Congestion Notification (ECN) marks received from capable hardware as a congestion signal.

DCTCP uses fraction of ECN marked packets to update congestion window. The window reduction ratio is always <= 1/2. Only when all of the packets are marked, congestion window is halved.

In order to keep the accuracy of the ECN marked fraction, a DCTCP receiver mirrors back incoming (or missing) CE marks by setting (or clearing) ECE marks. This feedback methodology is also adopted when the receiver uses delayed ACK.

The FreeBSD DCTCP implementation includes two minor modifications for the one-sided deployment. Considering the situation that DCTCP is used as sender and classic ECN is used as receiver, DCTCP sets the CWR flag as the reaction to the ECE flag. In addition, when classic ECN is used as sender and DCTCP is used as receiver, DCTCP avoids to mirror back ACKs only when the CWR flag is set in the incoming packet.

The other specifications are based on the paper and the RFC referenced in the *SEE ALSO* section below.

MIB Variables

The algorithm exposes the following tunable variables in the *net.inet.tcp.cc.dctcp* branch of the sysctl(3) MIB:

alpha The initial value to estimate the congestion on the link. The valid range is from 0 to 1024, where 1024 reduces the congestion window to half, if a CE is observed in the first window and *alpha* could not yet adjust to the congestion level on that path. Default is 1024.

shift_g An estimation gain in the *alpha* calculation. This influences the responsiveness when adjusting alpha to the most recent observed window. Valid range from 0 to 10, the default is 4, resulting in an effective gain of 1 / (2 ^ *shift_g*), or 1/16th.

slowstart A flag if the congestion window should be reduced by one half after slow start. Valid settings 0 and 1, default 0.

ect1 Controls if a DCTCP session should use IP ECT(0) marking when sending out segments (default), or ECT(1) marking making use of L4S infrastructure. Changes to this setting will

only affect new sessions, existing sessions will retain their previous marking value.

SEE ALSO

cc_cdg(4), cc_chd(4), cc_cubic(4), cc_hd(4), cc_htcp(4), cc_newreno(4), cc_vegas(4), mod_cc(4), tcp(4), mod_cc(9)

Mohammad Alizadeh, Albert Greenberg, David A. Maltz, Jitendra Padhye, Parveen Patel, Balaji Prabhakar, Sudipta Sengupta, and Murari Sridharan, "Data Center TCP (DCTCP)", *ACM SIGCOMM 2010*, http://research.microsoft.com/pubs/121386/dctcp-public.pdf, 63-74, July 2010.

Stephen Bensley, Dave Thaler, Praveen Balasubramanian, Lars Eggert, and Glenn Judd, *Data Center TCP (DCTCP): TCP Congestion Control for Data Centers*, https://tools.ietf.org/html/rfc8257.

HISTORY

The **cc_dctcp** congestion control module first appeared in FreeBSD 11.0.

The module was first released in 2014 by Midori Kato studying at Keio University, Japan.

AUTHORS

The **cc_dctcp** congestion control module and this manual page were written by Midori Kato *katoon@sfc.wide.ad.jp* and Lars Eggert *lars@netapp.com* with help and modifications from Hiren Panchasara *hiren@FreeBSD.org*