Measuring Multi-Path TCP Performance in Dual-Homed Devices

TECHNICAL REPORT

ICTEAM, Université catholique de Louvain (Belgium)

1 Introduction

This document describes in detail the metrics, methodology and testbed setup for the measurements on Multipath TCP performance. It is organized as follows. Section 2 describes the MPTCP Bandwidth and Delay Aggregation Benefit used for evaluating MPTCP performance with respect to TCP. Section 3 describes the architecture and methodology to perform MPTCP measurements. Section 4 describes the characteristics of the machines involved in the measurement testbed, and in particular the configuration of the server and the local and remote probes.

2 Metrics

The study of Multipath TCP performance is performed from the perspective of the final transport user perspective, and comparing basic performance aspects such as bandwidth (transport goodput) and file transfer delay in MPTCP with standard TCP. Two metrics are identified: a slight variation of the Bandwidth Aggregation Benefit metric, proposed by Kaspar (2011) [6] and already used in MPTCP analysis by Paasch (2013) [2], and the Delay Aggregation Benefit metric, derived from the same principle. Implementation of these metrics is detailed in section 3.

2.1 Transport Disjointivity

Given a set of n Internet paths $S = \{p_1, p_2, ..., p_n\}$, the transport disjointivity index (TDI) of S can be defined as the ratio between the maximum transport throughtput that can be achieved when using simultaneously all paths in S, denoted by C_{total}, and the addition of transport throughtput achievable (separately) over each path i, C_i:

$$TDI = \frac{C_{\text{total}}}{\sum_{i=1}^{n} C_i} \in [0, 1] \tag{1}$$

where maximum corresponds to the case in which $C_{\text{total}} = \sum_{i=1}^{n} C_i$, and which approaches minimum when the combined multipath transport throughput becomes negligible with respect to the addition of separate single-path transport throughputs ($C_{\text{total}} \ll \sum_{i=1}^{n} C_i$).

2.2 Bandwidth Aggregation Benefit

In this context, *bandwidth* between two endpoints is defined as the average goodput achievable during a certain amount of time between via a reliable transport protocol (TCP or MPTCP).

Let n be the number of networking interfaces available at the Measurement Agent (MA). Then, MPTCP will try to set up n different transport subflows (leading to potentially disjoint transport paths) between the MA and the Measurement Server (MS). For each MA's interface i, let C_i be the bandwidth capacity for the path traversed by the subflow associated to interface i and $C_{\max} = \max_{i \leq n} C_i$. Note that, if the corresponding path is compound by links $\{l_1, l_2, ..., l_m\}$, and $c(.)$ denotes the bandwidth capacity of a link, $C_i = \min_{i \leq m}\{c(l_i)\}$. Let C_{total} be the joint capacity of paths from every MA's interface i and the MS. Note that $C_{total} \leq \sum_{i=1}^{n} C_i$, equality holding when the n paths are completely disjoint (from a transport perspective, see section 2.1). This latter case of fully disjoint paths ($C_{total} = \sum_{i=1}^{n} C_i$) corresponds to the strict definition of Kaspar (2011) [6].

Then, the Bandwidth Aggregation Benefit when using MPTCP instead of regular TCP, and achieving a goodput x, can be defined as follows:

$$BW_{\text{benefit}}(x) = \begin{cases} \frac{x - C_{\max}}{C_{\text{total}} - C_{\max}}, & x \geq C_{\max} \\ \frac{x - C_{\max}}{C_{\max}}, & x < C_{\max} \end{cases} \in [-1, 1] \quad (2)$$

$BW = 1$ corresponds to the maximum benefit possible (MPTCP achieves all the available bandwidth C_{total}), $BW = -1$ corresponds to the worst case (MPTCP gets 0 Mbps), and $BW = 0$ implies no benefit, *i.e.*, MPTCP behaving as good as TCP.

2.3 Delay Aggregation Benefit

For a file to be transmitted between two endpoints by way of a reliable transport protocol (TCP or Multipath TCP), the term *delay* is defined as the amount of time between the transmission of the first bit and the reception of the acknowledgement for the last bit of the file. Note that this definition excludes the time required for the establishment of a TCP/MPTCP connection, which is considered separately in this work.

Then, let D_i be the delay when using a TCP connection over MA's interface i. $D_{\min} = \min_{i=1}^{n} D_i$, $D_{\max} = \max_{i=1}^{n} D_i$ and D^*_{th} defines the theoretical delay that could be ideally achieved if all available paths were used with perfect scheduling at the same time. In general, $D^*_{\text{th}} \leq D_{\min}$. Following the philosophy of Kaspar's aggregation benefit metrics, the Delay Aggregation Benefit when using MPTCP instead of regular TCP, and getting a delay x, can be defined within $[-1, 1]$ in a way such that it is:

- Positive if the delay with MPTCP is better (smaller) than the delay achieved by TCP over the fastest interface, and negative otherwise.

– Maximum when the MPTCP delay corresponds to (or outperforms) the best possible delay D^*_{th}.
– Minimum when MPTCP achieves a delay equal or larger than the TCP delay over the slowest interface, $D_{\max}$.

Or, equivalently:

$$D_{\text{benefit}}(x) = \begin{cases} 1 & , x \leq D^*_{\text{th}} \\ \frac{D_{\min}-x}{D_{\min}-D^*_{\text{th}}} & , D^*_{\text{th}} < x \leq D_{\min} \\ \frac{D_{\min}-x}{D_{\max}-D_{\min}} & , D_{\min} < x \leq D_{\max} \\ -1 & , x > D_{\max} \end{cases} \in [-1, 1] \tag{3}$$

In this paper, we explore the performance of Multipath TCP in a two-subflow scenario. In this context, the parameters taken into consideration into the Delay Benefit metric (Eq. (3)) can be computed as follows.

Let $k = \frac{D_{\max}}{D_{\min}}$ be a compression parameter. Then, it is immediate to observe that the smallest theoretical delay that MPTCP could reach in the considered two-subflow scenario is $\frac{k}{k+1}$ the minimum delay obtained in a single path.

Therefore, in the considered two-subflow scenario, the delay benefit is minimum (maximum penalty) when MPTCP delay is worse (longer) than the delay achieved by TCP over the slowest interface, and maximum when MPTCP delay is better (shorter) than $D^*_{\text{th}} = \frac{k}{k+1} D_{\min}$.

$$D^{\text{2-subflow}}_{\text{benefit}}(x) = \begin{cases} 1 & , x \leq \frac{k}{k+1} D_{\min} \\ (k+1)\frac{D_{\min}-x}{D_{\min}} & , \frac{k}{k+1} D_{\min} < x \leq D_{\min} \\ \frac{D_{\min}-x}{(k-1)D_{\min}} & , D_{\min} < x \leq kD_{\min} \\ -1 & , x > kD_{\min} \end{cases} \tag{4}$$

This characterization of $k = \frac{D_{\max}}{D_{\min}}$ allows to obtain a configuration-independent delay benefit of using MPTCP in scenarios with two subflows.

3 Measurement Architecture and Methodology

Experiments are performed by way of a specific software [7], designed to define and launch sequences of experiments between probes (Measurement Agents, MAs) and a centralized server (Measurement Server, MS). The MA determines the experiments setup and scheduling; measurements are reported to the MS. Architecture of this tool is displayed in the Fig. 1.

For each individual experiment launched by a MA, a control communication channel is set up between the MA and the MS and different measurements are collected over exchanges performed through one or several dedicated data communication channels.

The tool supports bandwidth and delay experiments, with two transmission modes emulating two standard patterns on user data transport: oneway transmission in MA $\longrightarrow$ MS direction (`oneway` mode), and request/response exchange in MA $\longleftarrow$ MS direction (`reqresp`). In each case, the delay is measured between

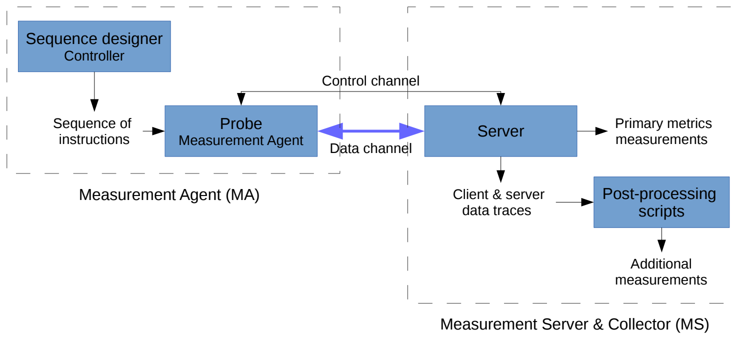

Fig. 1. Measurement tool architecture.

the transmission of MA's first bit until the the connection closing (by the MA after receiving the acknowledgement of the last bit from the MS, by the MS after receiving the acknowledgement of the last bit from the MA).

3.1 Procedure for an Individual Experiment

A typical experiment involves three main steps:

1. Hosts negotiate measurement parameters, over the control channel MA-MS.
2. The MA performs different active measurements, specific to the type of experiment, over one or several data channels MA-MS.
3. MA reports measurements to the MS, over the control channel.

MA reports involve the metrics computation, and can also include TCP and MPTCP connection traces. Traces report allow further and finer analysis by way of specific tools, such as *tcptrace* [1] and *mptcptrace* [5].

3.2 Types of Experiments

In bandwidth experiments, the MPTCP Bandwidth Aggregation Benefit is computed by measuring the amount of user data that can be sent over the available flows (TCP or MPTCP, see below) during a fixed amount of time (typically, 60 sec). In delay experiments, the MPTCP Delay Aggregation Benefit is computed by measuring the *delay* (in the sense of section 2.3) to transmit a fixed amount of bytes over the available flows (TCP or MPTCP, see below). Table 3.2 summarizes each experiment's main characteristics.

During an experiment, measurements on the network state are assumed to be valid for the duration of the experiment. That is, the tool implicitly assumes that the state of the network *does not change* during a single experiment, which can last several minutes.

A bandwidth experiment consists of the following active measurements:

Experiment	*Input*	*Mode*	*Main output*
Bandwidth	Transmission time (~ 60 sec)	`output|reqresp`	BW Aggregation Benefit $\in [-1, 1]$
Delay	Number of bytes $\in [10^3, 10^7]$ B	`output|reqresp`	Delay Aggregation Benefit $\in [-1, 1]$

Fig. 2. Main parameters of experiments.

1. For each available interface i, transmission over a data channel using TCP over the path associated to MA's interface i (for n MA's interfaces, n TCP MA $\longleftrightarrow$ MS connections).
2. Simultaneous TCP connections (data channels) over all paths associated to MA's available interfaces, to estimate the total available bandwidth for reliable transport (for n MA's interfaces, n TCP MA $\longleftrightarrow$ MS connections).
3. Transmission using MPTCP over a (multipath) data channel (1 MPTCP MA $\longleftrightarrow$ MS connection).

Comparing the total available TCP throughput obtained in steps 1 and 2 allows to estimate the transport disjointivity index (TDI) of paths between MA and MS, defined as in Eq. (1).

For a transmission time t, the expected time for the completion of a bandwidth experiment in the data plane is $(n+2) \times t$ (exchanges on the control plane have negligible duration with respect to active measurements in the data plane).

For a delay experiment, active measurements consists of:

1. For each available interface i, transmission over a data channel using TCP over the path associated to MA's interface i (for n MA's interfaces, n TCP MA $\longleftrightarrow$ MS connections).
2. Transmission using MPTCP over a data channel involving paths associated to all MA's available interfaces (1 MPTCP MA $\longleftrightarrow$ MS connection).

Given b bytes to be transferred, the expected time for the completion of a delay experiment in the data plane is thus $b \times \left(\sum_{i=1}^{n} \frac{1}{BW_{\mathrm{TCP}(i)}} + \frac{1}{BW_{\mathrm{MPTCP}}} \right)$.

In order to avoid possible bias in the results due to a particular order in the active measurements, the sequence of executed steps is randomized at each experiment.

3.3 Sequences of Experiments

In the measurements campaigns described in this paper, MAs execute sequences of independent experiments. These sequences are also *randomized* and stored in the MAs, for reproduceability purposes. Time between two consecutive experiments from the same MA can be constant (thus, leading to periodic MA measurements) or random following an exponential distribution (thus, emulating a Poisson event distribution on the MS), depending on the setup. See below for more details on the experimental setup (section 4).

4 Testbed Characteristics

A centralized Measurement Server (MS) is deployed in an Internet-reachable machine at UCL; measurement agents are expected to perform MPTCP performance measurements by running experiments against this MS. All involved machines use the MPTCP kernel implementation [4]. Main characteristics of the MS are detailed in Table 3. We have deployed a local MA at UCL and a remote MA at Aalto University (Finland).

	Server (MS)	UCL Probe (MA1)	Aalto Probe (MA2)
MPTCP version	0.88.8	0.88.x	0.89.2
Congestion control	LIA	OLIA, LIA, Cubic	Cubic
Receiving window	4096, 87380, 3752192	4096, 87380, 1009856	4096, 87380, 6291456
Sending window	4096, 16384, 3752192	4096, 16384, 1009856	4096, 16384, 4194304
Networking interfaces	`eth0`	`eth0.2`, `wlan1`	`eth0`, `eth1`, `wlan0`

Fig. 3. Main characteristics of the MS and the MAs at UCL (MA1) and Aalto (MA2). *Values for receiving and sending windows correspond to minimum, initial, maximum.*

4.1 Local Testbed (UCL)

For the local testbed experiments, the MA has been deployed on a router TL-WDR4900 with a MPTCP-capable kernel (see Table 3). This MA is connected to the Internet by way of two networking interfaces: an Ethernet interface attached to the UCL network, and a WiFi interface to a FON/Belgacom WiFi hotspot, able to reach the MS via a commercial ADSL connection. The two available paths between MA and MS are both topologically- and transport-disjoint. The path available through the MA wired interface shows a RTT in the order of milliseconds, with peaks at 40 msec. The path available through the MA wireless interface is significantly slower (between tens and hundreds of milliseconds, with peaks of seconds) and more variant. Fig. 4 shows schematically this setup.

The measurement sequences are computed and randomized at the MA, and the corresponding measurements are performed periodically, with a $5min$ ($300sec$) interval between the start of two consecutive experiments. For bandwidth experiments, each transport connection lasts $60sec$. For delay experiments, the amount of transmitted bytes is selected randomly, following a loguniform distribution between the corresponding bounds. The duration of each performed experiment is smaller than this value – in most cases substantially smaller.

Several measurement sequences have been performed in this scenario. Bandwidth configurations are parametrized by way of the Bandwidth Ratio (BWR), defined as the quotient between the goodput obtained on the MA's wired interface and the goodput on the MA's wireless (WiFi) interface. Scenarios with $BWR = \{1...15, 20\}$ are tested. Different BWR are emulated by shaping traffic

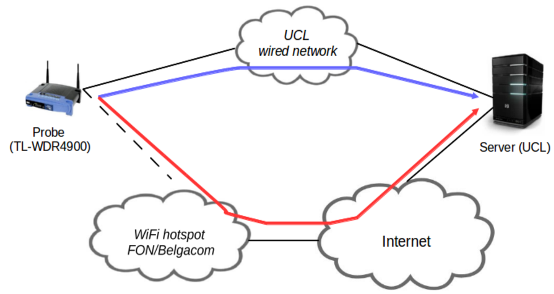

Fig. 4. Measurement setup at UCL.

in the MA wired (`eth`) interface, by way of the `tc` tool. The uploading rate on the wireless interface being ~ 1 Mbps, the traffic rate is shaped in the interface `eth0` at k Mbps, for $BWR = k$. Since traffic shaping only affects outgoing traffic (from the MA), only `oneway` experiments are performed. For each BWR configuration, bandwidth and delay measurements are performed. In the latter ones, the amount of bytes to be transmitted for each experiment is selected randomly in each sequence following a LogUniform distribution within intervals $[10^2, 10^6]$ bytes and $[10^3, 10^7]$ bytes. In total, 3064 experiments have been performed at disjoint time intervals; 1024 with OLIA configuration, 979 with LIA configuration and the rest with the MA default congestion control (Cubic). BWR configurations are distributed uniformly among the total number of experiments.

Fig. 5 displays the space of experiments on this scenario. Fig. 5(a) shows the time distribution of experiments (Y axis); the X axis indicates the amount of bytes transmitted, for delay experiments; and also represents bandwidth experiments. Fig. 5(b) presents the bandwidth configuration of performed experiments by way of the BWR parameter.

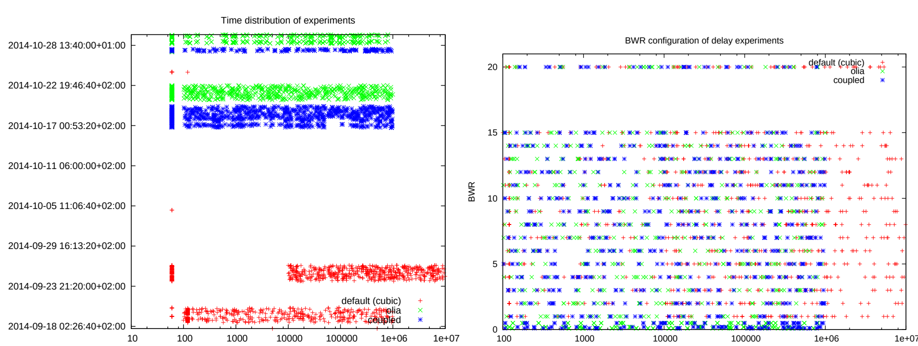

Fig. 5. Experiment space in the UCL testbed.

4.2 Remote MA configuration

The remote MA (MA2) runs a sequence of `oneway` and `reqresp` experiments. Time between two consecutive experiments is computed randomly, following an exponential distribution with $\lambda = 0.0023 \frac{\mathrm{exp}}{\mathrm{sec}}$, *i.e.*, 200 daily experiments in average. The MA has three networking interfaces (see Table 3). From our observations, paths between MA2 and MS in both directions are transport-disjoint, with estimated capacities (average) 4, 93 and 220 Mbps (MA2 $\longrightarrow$ MS) and 2, 22 and 50 Mbps (MS $\longrightarrow$ MA2). Path capacity estimations vary however significantly along the time. No traffic shaping is performed in this MA. In total, 132 experiments have been performed on this probe, half on each direction of communication.

Acknowledgements

This work has been partly supported by the EC's 7$^{\mathrm{th}}$ Framework Programme Grant No. 317647 (Leone), and by the Belgian IAP BESTCOM program. I am thankful to Pr. Olivier Bonaventure for his useful indications, and to Saba Ahsan, from Aalto University, for her support in the setup of remote experiments.

References

1. S. Ostermann: TCPTRACE, available from http://www.tcptrace.org.
2. C. Paasch, R. Khalili, O. Bonaventure: On the Benefits of Applying Experimental Design to Improve Multipath TCP. Proc. CoNEXT'13, Santa Bárbara, USA, Dec. 2013.
3. T. Dreibholz, M. Becke, H. Adhari, E. P. Rathgeb: Evaluation of A New Multipath Congestion Control Scheme using the NetPerfMeter Tool-Chain. Proc. SoftCOM'2011. Hvar, Croatia, Sept. 2011.
4. C. Paasch, S. Barré *et al.*: Multipath TCP in the Linux Kernel, available at http://www.multipath-tcp.org.
5. B. Hesmans, O. Bonaventure: Tracing Multipath TCP Connections. Proc. ACM SIGCOMM'14, Chicago, USA, 2014.
6. D. Kaspar: Multipath Aggregation of Heterogeneous Access Networks. PhD. 2011.
7. J. A. Cordero: MPTCP Bandwidth/Delay Benefit Measurements, available at http://inl.info.ucl.ac.be/softwares/mptcp-bwdelay-benefit-measurements.