ANALYSING INTERNET CONNECTION TRENDS IN DEVELOPING COUNTRIES IN RECENT TIMES

This work has been done by Ms.Bhagyashree Padalkar as a part of Outreachy Application for Measurement Lab

Introduction :
The Internet has revolutionized the computer and communications world like nothing before.The Internet is at once a world-wide broadcasting capability, a mechanism for information dissemination, and a medium for collaboration and interaction between individuals and their computers without regard for geographic location. The United Nations recently proposed that Internet access should be a human right. And yet only 1 out of 3 people in the world are connected via internet. Transmitting data—even a text message or a simple web page—requires bandwidth, something that's scarce in many parts of the world most of it being underdeveloped and developing countries in Asia and Africa. Through, this report I have tried to examine the recent trends in Internet Connectivity of developing countries (India). I have also further tried to compare these trends to their counterparts in developed countries (like UK, USA) to gain an overall perspective.

My Methodology :
I chose India to do this analysis due to the vast diversity of the country in terms of landscape,population,social and economic factors. It also one of the largest dataset of connections available amongst underdeveloped and developing countries in Asia and Africa.As a result we can study the trends in Internet Connectivity more effectively.

In this report I examine the some of the basic network connectivity measures of different connections from India. The network connectivity measures used are :
1.Average RTT
2.Download Throughput
3.% of Packets Retransmitted
4.% of times the connection is dropped

I have used median as aggregating measure since it is one of the least affected measures by big variations in data amongst all aggregating functions.It is hence one of the best representatives for the most occuring value for that function.

The data for connections from India was collected using Measurement Lab Dataset on Google Big Query over time period 2010-2012.

The data for connections from US,UK was collected using Measurement Lab Dataset on Google Big Query over the year 2012 as

we only want to establish a comparison over the most recent trends in developed countries.

While querying,

results from only NDT data were considered(project=0) as they were considered to be more specific and accurate for network diagnostics done.

Results from only server to client tests were considered as wanted to diagnose client-oriented values for network diagnostic tests (connection_spec.data_direction = 1)

Results were for only those data were considered when tests had been completed. As if not,then the test results cannot correctly estimate the connection performance
(

```
web100_log_entry.is_last_entry = True

web100_log_entry.snap.HCThruOctetsAcked >= 8192

(web100_log_entry.snap.SndLimTimeRwin +
 web100_log_entry.snap.SndLimTimeCwnd +
 web100_log_entry.snap.SndLimTimeSnd) >= 9000000

(web100_log_entry.snap.SndLimTimeRwin +
web100_log_entry.snap.SndLimTimeCwnd +
web100_log_entry.snap.SndLimTimeSnd) < 3600000000

(web100_log_entry.snap.State == 1
        OR (web100_log_entry.snap.State >= 5
            AND web100_log_entry.snap.State <= 11))
```

)

Queries used to extract and process the data were as follows :

Average RTT of connections from the country :

The following query extracts the median of Average RTT values of the country(here India) over a period of one month(here Mar 2010)

The condition to get Average RTT values is :
web100_log_entry.snap.SumRTT/web100_log_entry.snap.CountRTT
We have further averaged it over IPs before calculating the median for the country as a whole.

Here it makes sense to exclude results of tests with fewer than 10 round trip time samples, because there are not enough samples to accurately estimate Average RTT

```
(web100_log_entry.snap.CountRTT > 10)
```

Final Query :

```
SELECT percentile_cont(0.5) OVER (ORDER BY rtt)
FROM (
SELECT
web100_log_entry.connection_spec.remote_ip AS ips,
AVG(web100_log_entry.snap.SumRTT/web100_log_entry.snap.CountRTT)
AS rtt
FROM [measurement-lab:m_lab.2010_03]
WHERE
IS_EXPLICITLY_DEFINED(web100_log_entry.connection_spec.remote_ip)
AND
IS_EXPLICITLY_DEFINED(connection_spec.client_geolocation.country_n
ame)
AND connection_spec.client_geolocation.country_name='India'
AND IS_EXPLICITLY_DEFINED(web100_log_entry.log_time)
AND web100_log_entry.log_time > PARSE_UTC_USEC('2010-03-01
00:00:00') / POW(10, 6)
AND web100_log_entry.log_time < PARSE_UTC_USEC('2010-03-31
23:59:59') / POW(10, 6)
AND
IS_EXPLICITLY_DEFINED(web100_log_entry.connection_spec.local_ip)
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.HCThruOctetsAcked)
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeRwin)
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeCwnd)
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeSnd)
AND IS_EXPLICITLY_DEFINED(project)
AND project = 0
AND IS_EXPLICITLY_DEFINED(connection_spec.data_direction)
AND connection_spec.data_direction = 1
AND IS_EXPLICITLY_DEFINED(web100_log_entry.is_last_entry)
AND web100_log_entry.is_last_entry = True
AND web100_log_entry.snap.HCThruOctetsAcked >= 8192
AND (web100_log_entry.snap.SndLimTimeRwin +
     web100_log_entry.snap.SndLimTimeCwnd +
     web100_log_entry.snap.SndLimTimeSnd) >= 9000000
AND (web100_log_entry.snap.SndLimTimeRwin +
     web100_log_entry.snap.SndLimTimeCwnd +
     web100_log_entry.snap.SndLimTimeSnd) < 3600000000
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.MinRTT)
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SumRTT)
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.CountRTT)
AND web100_log_entry.snap.CountRTT > 10
AND (web100_log_entry.snap.State == 1
        OR (web100_log_entry.snap.State >= 5
            AND web100_log_entry.snap.State <= 11))
GROUP BY ips )
```

Download Throughput of connections from the country :

The following query extracts the median of Download Throughput values of the country(here India) over a period of one month(here Nov 2010)

The condition to get Download Throughput values is :
web100_log_entry.snap.HCThruOctetsAcked/
(web100_log_entry.snap.SndLimTimeRwin +
web100_log_entry.snap.SndLimTimeCwnd +
web100_log_entry.snap.SndLimTimeSnd)
We have further averaged it over IPs before calculating the median for the country as a whole.

Also we havent considered those NDT tests where
web100_log_entry.snap.CongSignals = 0
i.e. The test ends during slow start and never reaches congestion. As this happens if the test was interrupted by the user or due to some errors.

Final Query :

```
SELECT percentile_cont(0.5) OVER (ORDER BY thru)
FROM (SELECT web100_log_entry.connection_spec.remote_ip AS ips ,
AVG(web100_log_entry.snap.HCThruOctetsAcked/
(web100_log_entry.snap.SndLimTimeRwin +
web100_log_entry.snap.SndLimTimeCwnd +
web100_log_entry.snap.SndLimTimeSnd)) AS thru,
FROM [measurement-lab:m_lab.2010_11] WHERE

IS_EXPLICITLY_DEFINED(web100_log_entry.connection_spec.remote_ip)
   AND
IS_EXPLICITLY_DEFINED(connection_spec.client_geolocation.country_n
ame)
   AND connection_spec.client_geolocation.country_name='India'
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.log_time)
   AND web100_log_entry.log_time > PARSE_UTC_USEC('2010-11-01
00:00:00') / POW(10, 6)
   AND web100_log_entry.log_time < PARSE_UTC_USEC('2010-11-30
23:59:59') / POW(10, 6)
   AND
IS_EXPLICITLY_DEFINED(web100_log_entry.connection_spec.local_ip)
   AND
IS_EXPLICITLY_DEFINED(web100_log_entry.snap.HCThruOctetsAcked)
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeRwin)
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeCwnd)
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeSnd)
   AND IS_EXPLICITLY_DEFINED(project)
```

```
   AND project = 0
   AND IS_EXPLICITLY_DEFINED(connection_spec.data_direction)
   AND connection_spec.data_direction = 1
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.is_last_entry)
   AND web100_log_entry.is_last_entry = True
   AND web100_log_entry.snap.HCThruOctetsAcked >= 8192
   AND (web100_log_entry.snap.SndLimTimeRwin +
        web100_log_entry.snap.SndLimTimeCwnd +
        web100_log_entry.snap.SndLimTimeSnd) >= 9000000
   AND (web100_log_entry.snap.SndLimTimeRwin +
        web100_log_entry.snap.SndLimTimeCwnd +
        web100_log_entry.snap.SndLimTimeSnd) < 3600000000
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.CongSignals)
   AND web100_log_entry.snap.CongSignals > 0
   AND (web100_log_entry.snap.State == 1
        OR (web100_log_entry.snap.State >= 5
            AND web100_log_entry.snap.State <= 11))
   GROUP BY ips

  );
```

3. % of packets retransmitted

The following query extracts the median % of packets retransmitted values of the country(here India) over a period of one month(here Nov 2010)

The condition to get packets retransmitted is : (web100_log_entry.snap.SegsRetrans/web100_log_entry.snap.DataSegsOut)
We have further averaged it over IPs before calculating the median for the country as a whole.

To convert to percentage I have multiplied the final value by 100

Final Query :

```
SELECT percentile_cont(0.5) OVER (ORDER BY retrans)
FROM (SELECT web100_log_entry.connection_spec.remote_ip AS ips,
AVG(web100_log_entry.snap.SegsRetrans/web100_log_entry.snap.DataSegsOut) AS retrans
FROM [measurement-lab:m_lab.2010_11] WHERE

IS_EXPLICITLY_DEFINED(web100_log_entry.connection_spec.remote_ip)
   AND
IS_EXPLICITLY_DEFINED(connection_spec.client_geolocation.country_name)
   AND connection_spec.client_geolocation.country_name='India'
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.log_time)
   AND web100_log_entry.log_time > PARSE_UTC_USEC('2010-11-01 00:00:00') / POW(10, 6)
   AND web100_log_entry.log_time < PARSE_UTC_USEC('2010-11-30
```

```
23:59:59') / POW(10, 6)
   AND
IS_EXPLICITLY_DEFINED(web100_log_entry.connection_spec.local_ip)
   AND
IS_EXPLICITLY_DEFINED(web100_log_entry.snap.HCThruOctetsAcked)
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeRwin)
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeCwnd)
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeSnd)
   AND IS_EXPLICITLY_DEFINED(project)
   AND project = 0
   AND IS_EXPLICITLY_DEFINED(connection_spec.data_direction)
   AND connection_spec.data_direction = 1
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.is_last_entry)
   AND web100_log_entry.is_last_entry = True
   AND web100_log_entry.snap.HCThruOctetsAcked >= 8192
   AND (web100_log_entry.snap.SndLimTimeRwin +
        web100_log_entry.snap.SndLimTimeCwnd +
        web100_log_entry.snap.SndLimTimeSnd) >= 9000000
   AND (web100_log_entry.snap.SndLimTimeRwin +
        web100_log_entry.snap.SndLimTimeCwnd +
        web100_log_entry.snap.SndLimTimeSnd) < 3600000000
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SegsRetrans)
   AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.DataSegsOut)
   AND web100_log_entry.snap.DataSegsOut > 0
   AND (web100_log_entry.snap.State == 1
        OR (web100_log_entry.snap.State >= 5
          AND web100_log_entry.snap.State <= 11))
  GROUP BY ips
     );
```

4. % of times the connection is dropped

If web100_log_entry.snap.MinRTT = 4294967295(this is the upper base limit for the variable) we can say that the MinRTT for the test was 4294967295 ms and thus the connection cannot be established in real time and was dropped

To compute the number of connections from the country(here India) dropped over the time period(here Apr 2012), we use the following query

```
SELECT COUNT(ips)
FROM (SELECT web100_log_entry.connection_spec.remote_ip AS ips
FROM [measurement-lab:m_lab.2012_04]
WHERE
IS_EXPLICITLY_DEFINED(connection_spec.client_geolocation.country_n
ame)
AND connection_spec.client_geolocation.country_name='India'
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.MinRTT)
AND  web100_log_entry.snap.MinRTT = 4294967295
AND IS_EXPLICITLY_DEFINED(web100_log_entry.log_time)
AND web100_log_entry.log_time > PARSE_UTC_USEC('2012-04-01
00:00:00') / POW(10, 6)
```

```
AND web100_log_entry.log_time < PARSE_UTC_USEC('2012-04-30
23:59:59') / POW(10, 6)
GROUP BY ips);
```

To compute the total number of connections from the country(here India) over the time period(here Apr 2012), we use the following query

```
SELECT COUNT(ips)
FROM (SELECT web100_log_entry.connection_spec.remote_ip AS ips
FROM [measurement-lab:m_lab.2012_04]
WHERE
IS_EXPLICITLY_DEFINED(connection_spec.client_geolocation.country_n
ame)
AND connection_spec.client_geolocation.country_name='India'
AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.MinRTT)
AND IS_EXPLICITLY_DEFINED(web100_log_entry.log_time)
AND web100_log_entry.log_time > PARSE_UTC_USEC('2012-04-01
00:00:00') / POW(10, 6)
AND web100_log_entry.log_time < PARSE_UTC_USEC('2012-04-30
23:59:59') / POW(10, 6)
GROUP BY ips);
```

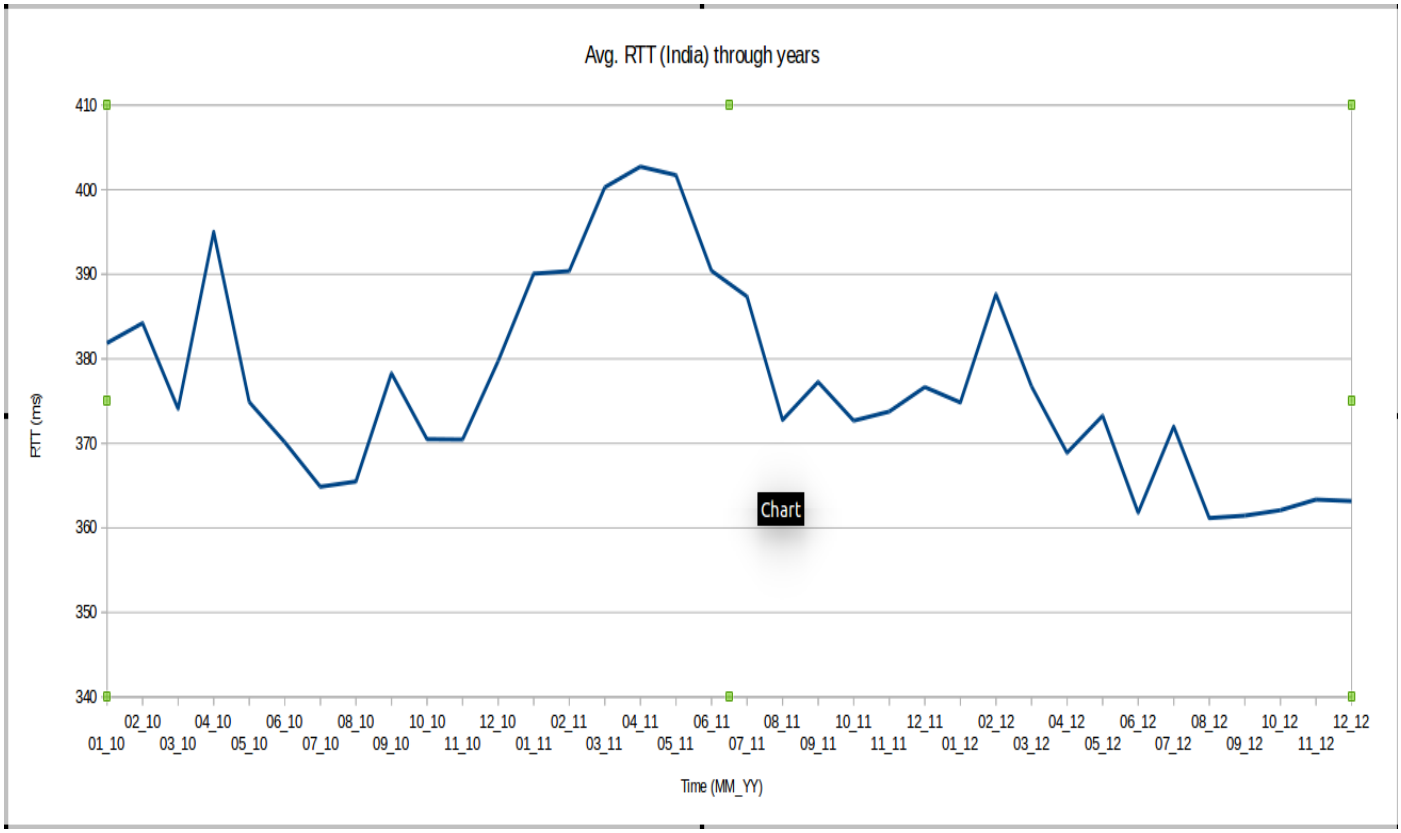
Divding the first value by second for the same time period will give us the fraction of connections dropped over that time period. Multiplying this value by 100 will give us the required %

Results :

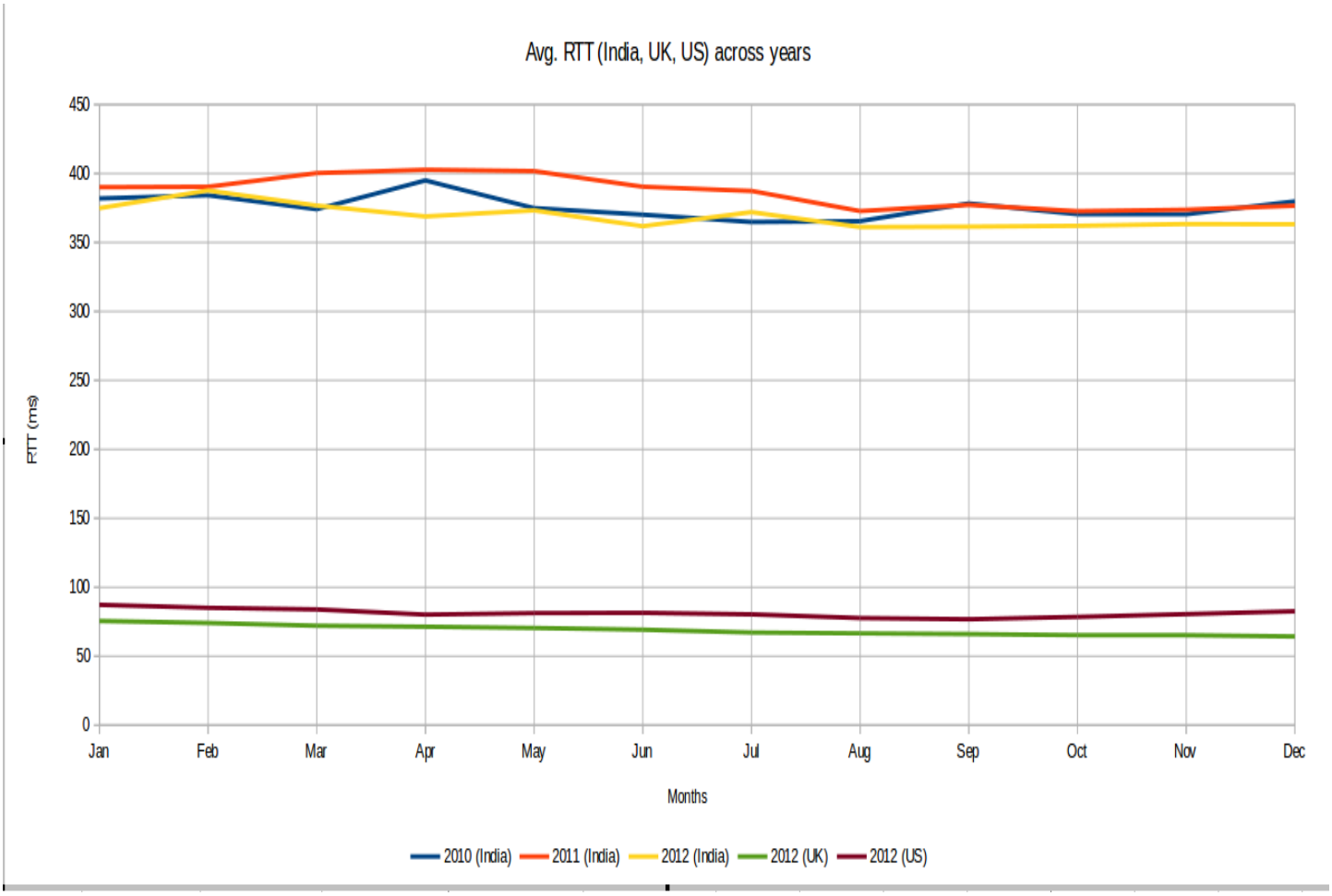
As we can see fromthe graphs the values of the parameters in the latter half of 2010 and 2011 intermingle but connections in India have become significantly better in 2012 as characterized by lower Avg RTT values,Higher Download Throughput,Lower % of packets retransmitted and lower % of packets dropped. This fact can be attributed to the availability of 3G data connections to the public since 2012 and the subsequent bandwidth increase.

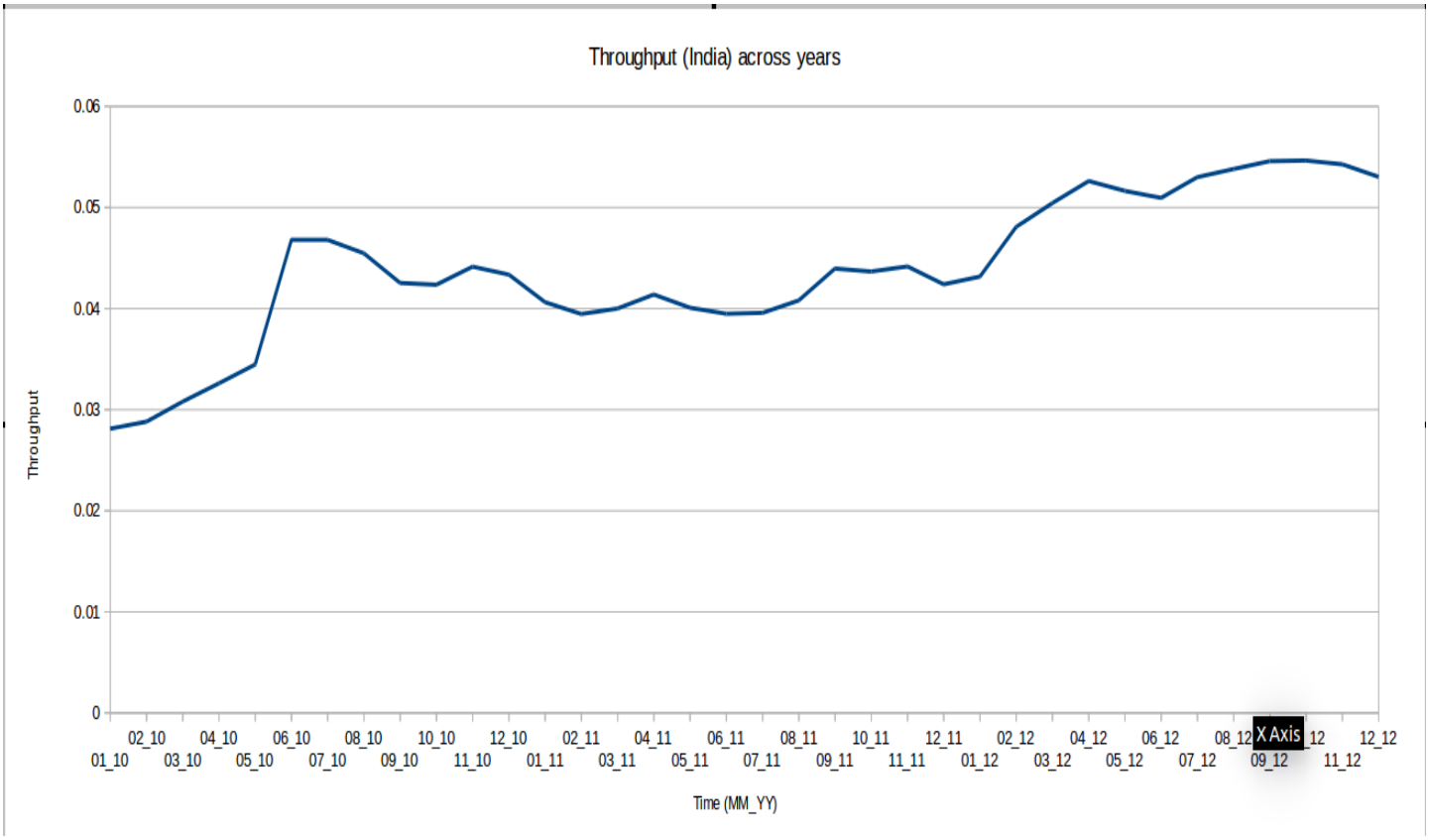
Also the earlier half of 2010 can be characterized by low Avg RTT values,low Download Throughput,Low % of packets retransmitted and high % of packets dropped.While these are conflicting results in itself, we leave it for further inspection whether this is an anamoly.

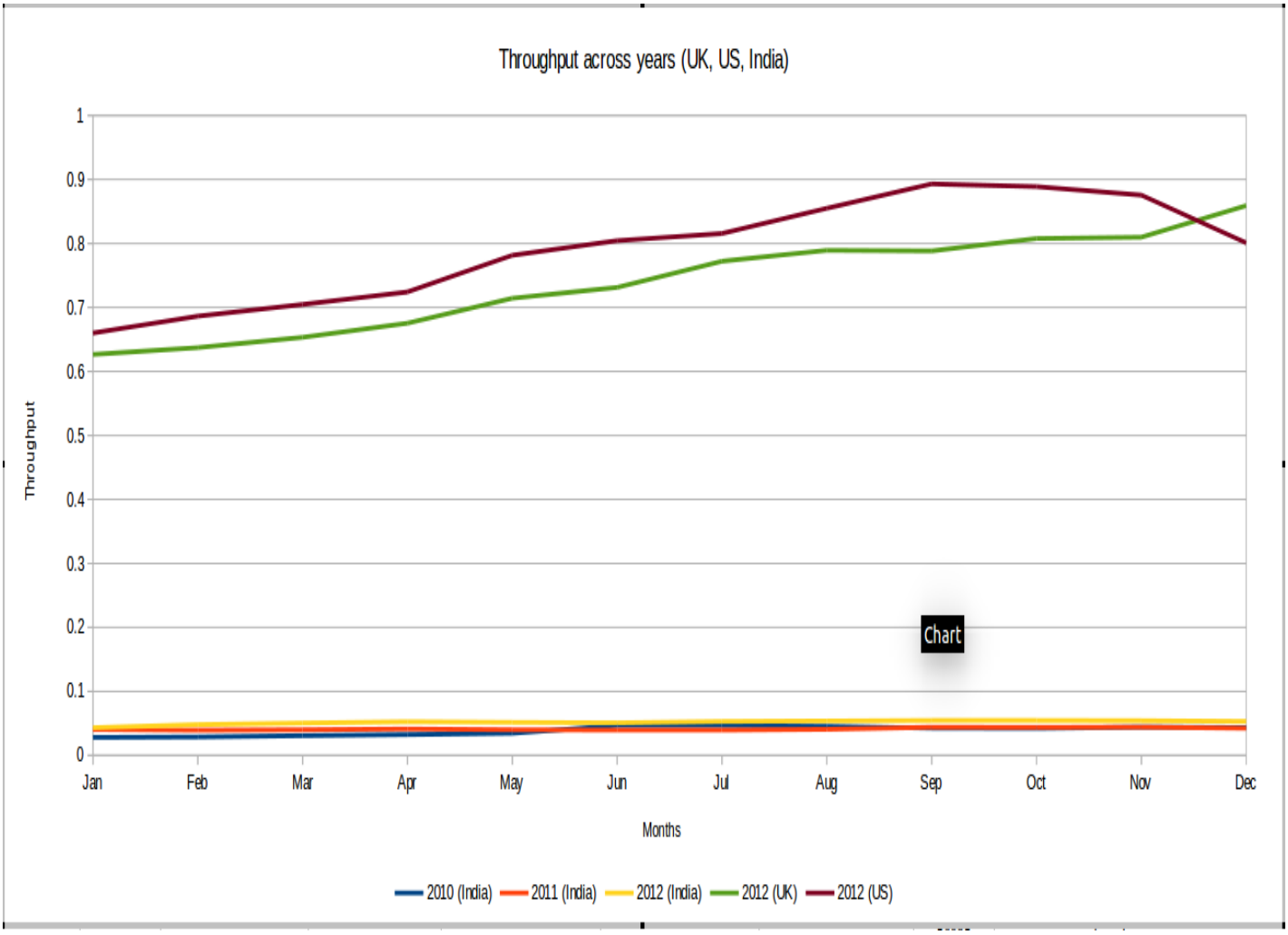
Also,as we can see from the comparison graphs for the same parameters against developed countries like US and UK, the connections in India are lagging very behind. Connections in India are almost 5 times slower, Download Throughput 13 times less, % of packets needed to be retransmitted almost 18 times more.

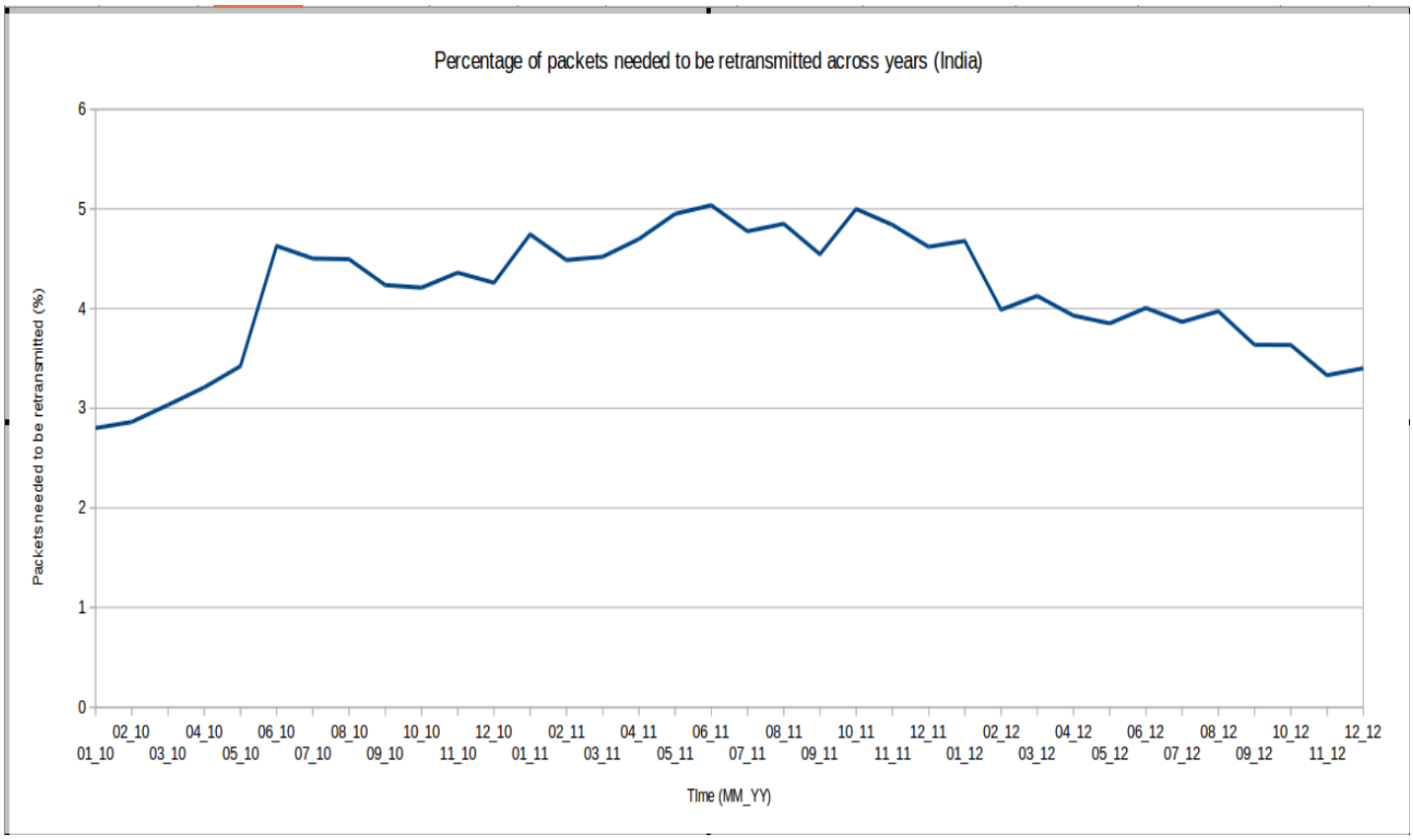


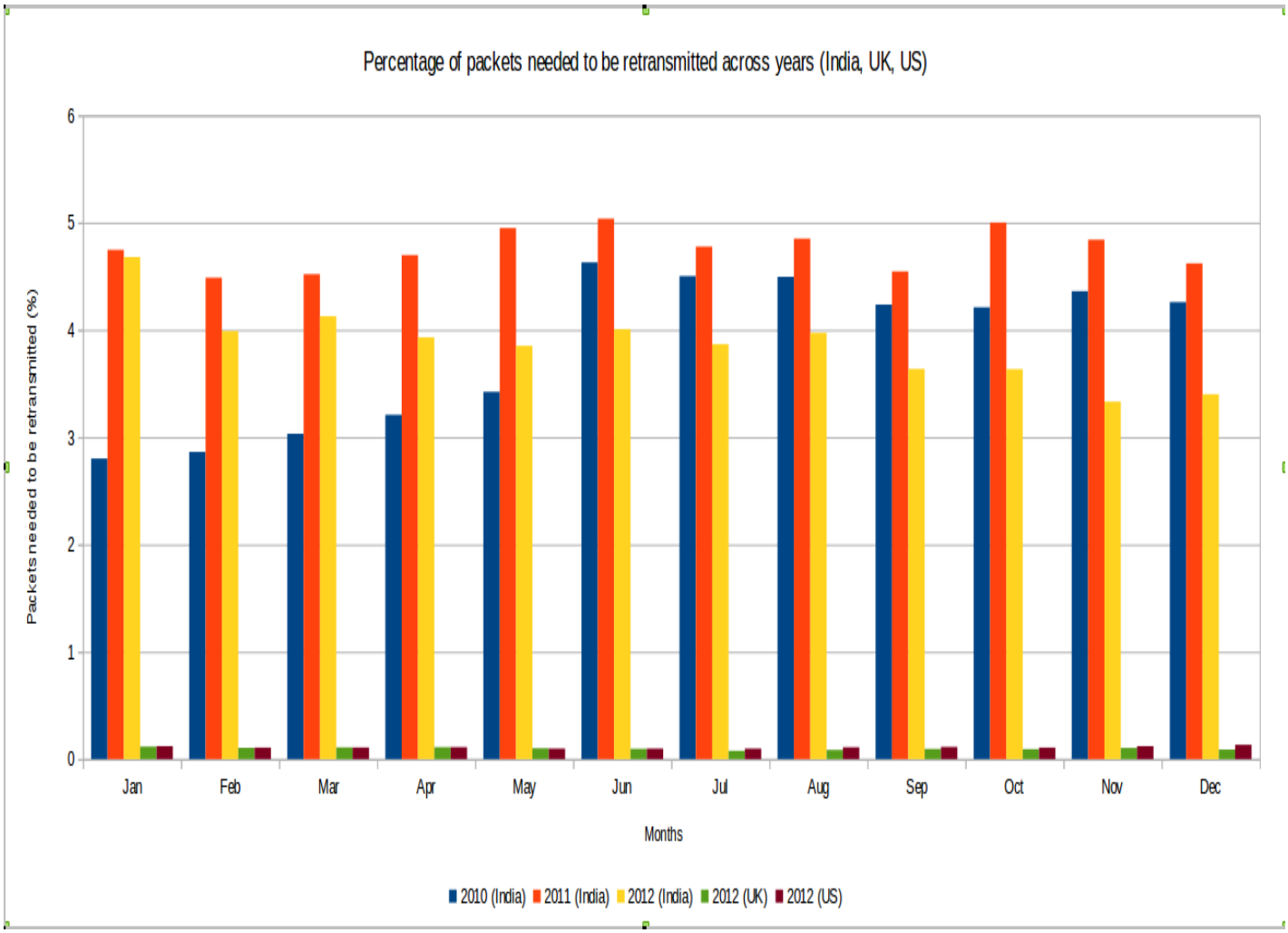
Average RTT in ms and Download Throughput in Mbps.

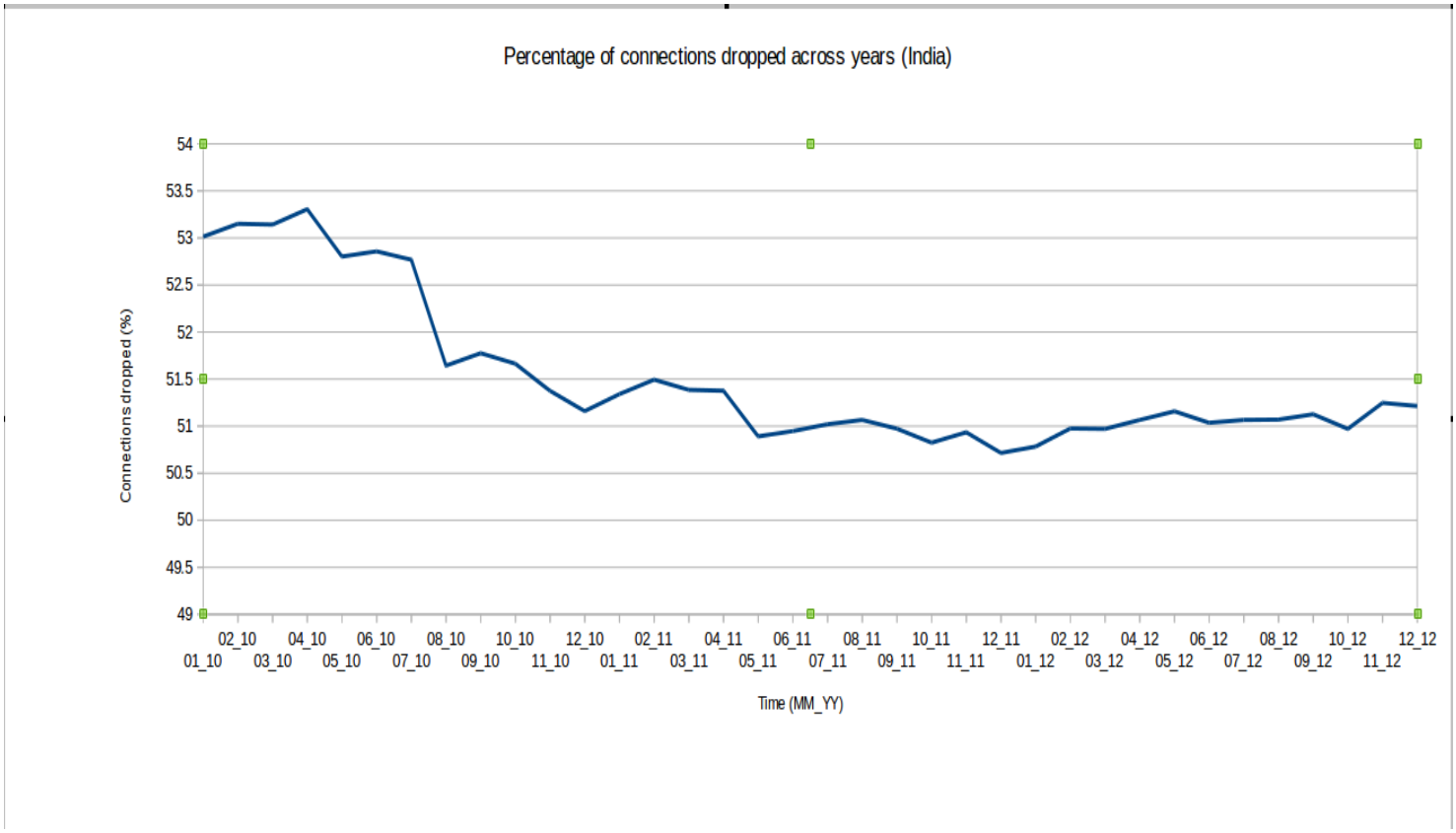


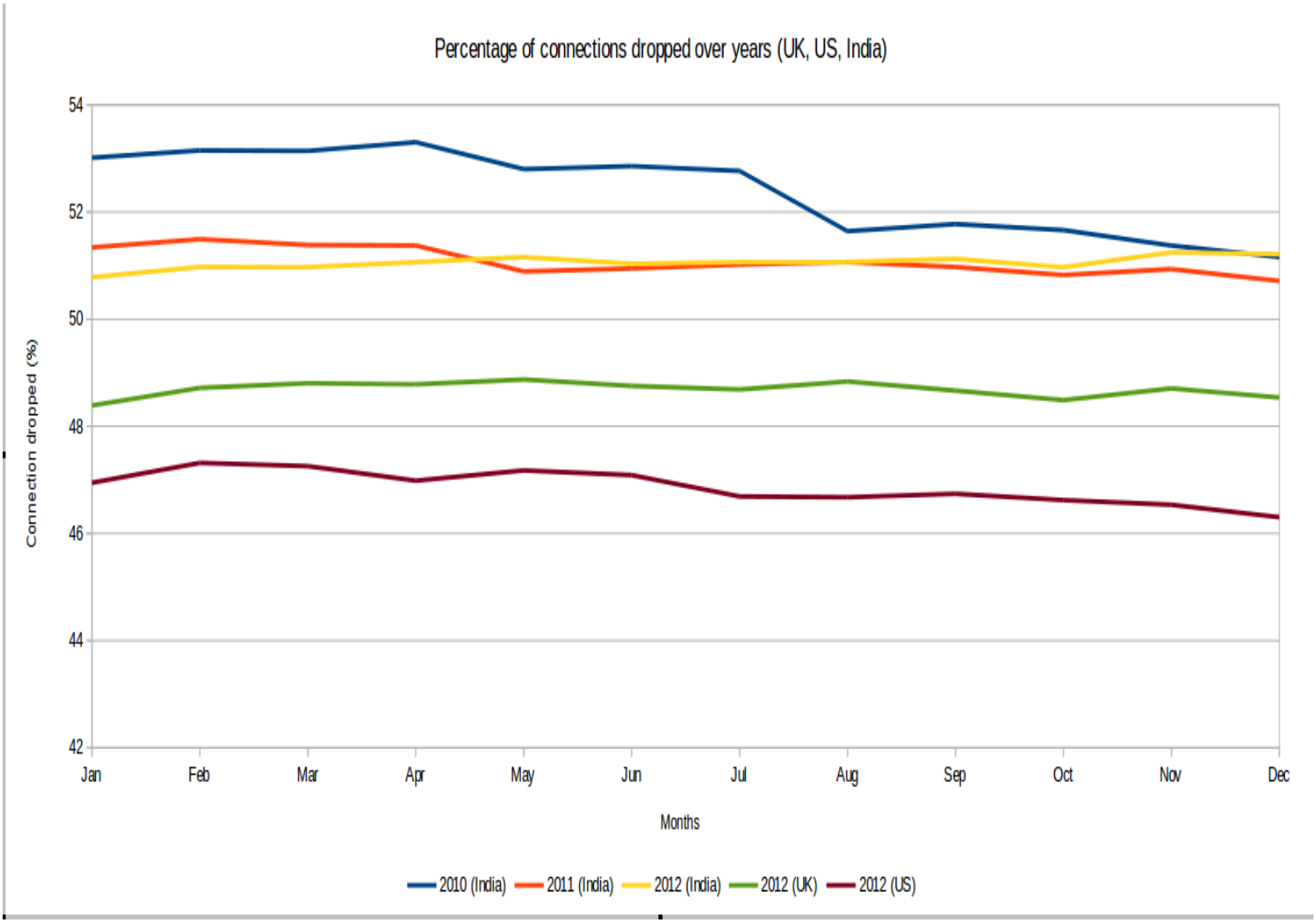

Throughput (India) across years
Throughput
0.06
0.05
0.04
0.03
0.02
0.01
0
02_10 04_10 06_10 08_10 10_10 12_10 02_11 04_11 06_11 08_11 10_11 12_11 02_12 04_12 06_12 08_12 10_12 12_12
01_10 03_10 05_10 07_10 09_10 11_10 01_11 03_11 05_11 07_11 09_11 11_11 01_12 03_12 05_12 07_12 09_12 11_12
X Axis
Time (MM_YY)


Throughput across years (UK, US, India)
Throughput
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Chart
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Months
2010 (India)
2011 (India)
2012 (India)
2012 (UK)
2012 (US)


Percentage of packets needed to be retransmitted across years (India)
Packets needed to be retransmitted (%)
0
1
2
3
4
5
6
01_10
02_10
03_10
04_10
05_10
06_10
07_10
08_10
09_10
10_10
11_10
12_10
01_11
02_11
03_11
04_11
05_11
06_11
07_11
08_11
09_11
10_11
11_11
12_11
01_12
02_12
03_12
04_12
05_12
06_12
07_12
08_12
09_12
10_12
11_12
12_12
Time (MM_YY)


Percentage of packets needed to be retransmitted across years (India, UK, US)
Packets needed to be retransmitted (%)
0
1
2
3
4
5
6
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2010 (India)
2011 (India)
2012 (India)
2012 (UK)
2012 (US)


Percentage of connections dropped across years (India)
Connections dropped (%)
54
53.5
53
52.5
52
51.5
51
50.5
50
49.5
49
02_10 04_10 06_10 08_10 10_10 12_10 02_11 04_11 06_11 08_11 10_11 12_11 02_12 04_12 06_12 08_12 10_12 12_12
01_10 03_10 05_10 07_10 09_10 11_10 01_11 03_11 05_11 07_11 09_11 11_11 01_12 03_12 05_12 07_12 09_12 11_12
Time (MM_YY)


Percentage of connections dropped over years (UK, US, India)
Connection dropped (%)
54
52
50
48
46
44
42
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Months
2010 (India)
2011 (India)
2012 (India)
2012 (UK)
2012 (US)