

Multipath IP transmission: Motivation, Design, Performance

ABSTRACT

Multipath Internet transmission has become available in recent years. Most devices have more than one interface that can connect to Internet. Especially the booming of mobile devices introduces the greatest chance for multipath communication.

In 2011, multipath TCP was introduced as RFC by IETF. This draft defines the architecture and guideline of multipath TCP development. Also, a popular multipath TCP implementation[1] attracts active research on multipath based on this implementation. But as the name refers, multipath TCP is only designed for TCP traffic. Inside the current Internet infrastructure, it is still an accepted assumption that most Internet traffic is transmitted via the TCP protocol. However, the rise of new streaming applications and P2P protocols that try to avoid traffic shaping techniques will likely increase the use of other protocols like UDP as a transport protocol.

In this paper, we introduce multipath communication at network layer instead of transmission layer. By introducing multipath capability at network layer, almost all traffic on Internet can benefit from multipath. Also, in the kernel of any operation system, because the huge difference of complexity between TCP and IP, we expect a much lighter weighted implementation of multipath at network layer. To evaluate the performance of MPIP, we implement our proposition in the latest Linux kernel under Ubuntu system. We prove that besides having the same even better performance as MPTCP for TCP traffic, MPIP also has good support to UDP traffic, and provides potential optimized routing decision for specific applications.

1. INTRODUCTION

Multipath has become available in recent years. Also, IETF proposed RFC 6182 specifically for multipath TCP in 2011. By introducing multipath, not only higher throughput can be achieved, the reliability of connections also benefits because multipath on one connection can achieve failover, i.e., if one path fails, the others will keep working without influencing applications.

Most current devices (Mainly mobile devices) have more than one internet interface (4G, WiFi), it is possible to make use of this facility to improve the quality of Internet trans-

mission. In scenarios that end users want high throughput, parallel multipath transmission can possibly improve throughput. In scenarios that end users have intermittent internet connection on one interface, multipath connection can provide smooth switching between connections.

Current work on multipath is mainly on TCP. In MPTCP, if the user has more than one Internet interface, there will be more than one subflow in one TCP connection while each subflow is an independent TCP connection. In this way, the user does not need to re-establish the connection when switching connection. But MPTCP can only be used in TCP connection given that there is still large amount of non-TCP traffic on the Internet although TCP traffic is dominating.

To enable non-TCP traffic to use multipath feature, we propose our multipath implementation at network layer. Introducing multipath at network layer has several benefits over transportation layer. At transportation layer, the complexity of TCP protocol makes the design of multipath over TCP more complicated and vulnerable. These built-in characters of MPTCP predesignate that MPTCP is not flexible enough to satisfy various requirement of different applications. As for MPIP, the simplicity of IP protocol makes it much easier to implement. And almost all traffic on the Internet will go through IP protocol, this makes MPIP more omnipotent than MPTCP. As will show in later sections, because of heterogeneous characters of different NICs on one device, different configuration can generate totally different user experience. If MPIP can provide equivalent performance with MPTCP for TCP traffic, then the good flexibility of MPIP definitely provides a much better option for multipath deployment.

By implementing the prototype of MPIP, our contribution is three-fold.

1. We propose the overall design and architecture of MPIP. By comparing our design with MPTCP, we see that implementing multipath at network layer has much lighter weight and more straightforward than transportation layer.
2. We implement our design in the latest Linux kernel under Ubuntu system. Also, we evaluate the implementation in different Internet environments. We show that our implementation can match MPTCP in TCP protocol, and also, other protocols like UDP can fit perfectly

with multipath IP.

3. For investigation purpose, we combine the implementation of MPTCP and MPIP together to prove multipath feature at both layers. It turns out that this combination can provide better and more consistent performance over certain Internet conditions.

The rest of the paper is organized as follows. Section 2 describes the related work. The design of our implementation is introduced in Section 3. In Section 4, we report the experimental results for our multipath IP design. We conclude the paper with summary and future work in Section 5.

2. RELATED WORK

Multipath transmission has been attracting research interest for quite some time. Back to 2001, Hsieh et al proposed pTCP[2] that effectively performs bandwidth aggregation on multi homed mobile hosts with simulation results. In [3], the authors investigated the potential benefits of coordinated congestion control for multipath data transfers as well as relative path selection algorithm when multipath is available. In [4], Dong et al implemented concurrent TCP(cTCP) in FreeBSD to improve throughput of connections through balancing traffic load on multiple end-to-end paths. The Stream Control Transmission Protocol (SCTP)[5] is designed with multihoming to support failover instead of parallel transmission. In [6], Joe et al extends SCTP to support simultaneous transmission with best load-sharing and achieve throughput improvement.

In 2010, Barre et al published an experimental result of using multiple paths simultaneously in TCP transmission[7]. After two years, the same team released the first public implementation of multipath TCP in [1]. In [1], according to RFC 6182 released by IETF specifically for multipath TCP in 2011, the authors implemented a complete prototype of multipath TCP in Linux and Android system. They also explored many other aspects of MPTCP in [8], [9], [10], [11].

Base on this prototype, many researches that focus on measurement and improvement of MPTCP join the research of multipath. In [12], based on WiFi and cellular network Chen et al did a thorough measurement of MPTCP over wireless environment. They considered multiple cellular network providers to do complete side-by-side comparison. In [13], a delay-based congestion control algorithm which is a transformation of Vegas[14] was proposed for multipath TCP.

At router level, ECMP[15] supports load balancing of general IP packets among different paths between two routers. But ECMP need router's support and doesn't work on end hosts.

In industry, Apple[16] implements multipath TCP in iOS 7.0 for its siri, cloud-based natural language voice command and navigation service which is the first large-scale deployment of multipath TCP.

Table 1: Control Message Structure

Control Message
<i>Node ID,</i>
<i>Session ID,</i>
<i>Path ID,</i>
<i>Feedback Path ID,</i>
<i>Packet Timestamp,</i>
<i>Path Delay</i>
<i>Local Address List</i>
<i>CM Flags</i>
<i>Checksum</i>

3. THE DESIGN

Based on Linux kernel 3.12, the implementation is mainly at the network layer and targets to IP protocol. To make the implementation transparent to users and other layers, we try to keep all kernel modifications independent from other modules. To keep the simplicity of IP protocol, we keep the connectionless feature of IP protocol while maintaining some feedback information of different paths. As will be shown later, we achieve this goal by simply keeping track of several tables. This can make sure that we don't add excessive overhead to the system.

As a fundamental problem, NAT devices are commonly used in the current Internet. For a single connection, multiple NAT devices can exist on the path. With the existence of NAT devices, many IP addresses are exposed to the Internet as a combination of IP address and port number. So in the following sections, when referring the address of a node, we will use the combination of IP address and port number.

3.1 IP Layer Control Communication

In TCP protocol, the ACK packet is mainly used as the feedback information from the receiver. Then the sender is able to know the real time status of the transmission. But IP protocol doesn't have this built-in feedback loop because it is a connectionless protocol. To maintain multiple paths at IP layer, we need to add control communication functionality like the ACK packet in TCP. Instead of constructing new control packets between connection ends, we use piggyback technology to implement the feedback message in IP layer. This can avoid excessive control packets and reduce overhead of MPIP.

For each MPIP enabled packet that goes out of the system, we add an additional control message(CM) data block at the end of user data. Sometimes the size of the packet will exceeds MTU after attaching the data block. In this case, we reduce the amount of user data from transportation layer to fit the CM block, e.g, we decrease the MSS value for TCP connection.

The content in the control message is shown in Table 1.

Node ID is the globally unique identification of one node. The combination of IP address and port number is not a qual-

ified candidate because during connection switching, IP address can change on a specific NIC. To have a static node ID, we choose the MAC address of any one NIC on the node to be the unique ID. The value of node ID is initiated when the system starts and keep it unchanged until the system exits. Every time the node sends out a packet, it fills the field of *node ID* with this value into the control message.

Local Address List carries all local IP addresses. This list will be used to construct new MPIP paths.

CM Flags notates the functionality of the packet. With different value of *CM Flags*, different action will be taken when the packet is received.

Checksum is used to verify the validation of the CM data block. This value is assigned by simply adding up the value of all other field in the CM data block. This will be recalculated when received to judge whether a CM data block is attached in this packet. The packet will be treated as a normal packet if the checksum verification doesn't get through.

Other fields of the CM block will be explained in following sections.

3.2 MPIP availability handshake

As a new feature in Linux, MPIP needs to be backward compatible. For one specific connection, before enabling MPIP at both end, the two sides need to synchronize with each other to make sure both are MPIP enabled. Locally, every node maintains Table 2 to identify the availability of MPIP for its opposite nodes. Before the handshake finishes, all communication on the connection is normal traffic without MPIP enabled.

Table 2: MPIP availability

IP Address	Port Number	MPIP Availability	Query Count
IP_1	P_1	True	2
IP_2	P_2	False	5

When a node sends out a packet, it checks locally whether the target node is MPIP enabled in Table 2. If not, besides sending out the original packet, the system copies the current packet and inserts the CM block into the copied packet with *CM Flags* of *Flags_Enable*. This value is used for MPIP query. When this packet is received by a MPIP enabled node, the receiver adds the sender's IP address and port number into Table 2 with value of *True*, then sends back the confirmation to the sender with *CM Flags* of *Flags_Enabled* at a proper time. For TCP and non-TCP connection, we send out the confirmation packet at different times. For any protocol rather than TCP, we populate a new packet, fill all header fields with the correct information, and attach the CM block at the end, then send back to the sender right away. But for TCP, we add the confirmation request into a waiting list, and piggyback the confirmation when next TCP message is sent out to that specific node. There are two reasons to have this different process.

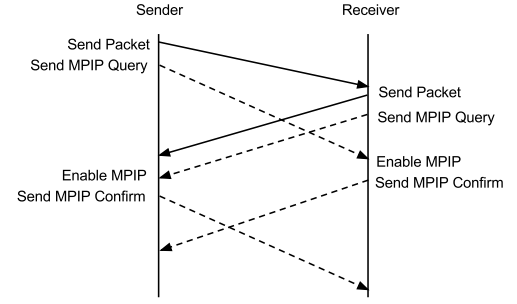


Figure 1: MPIP Handshake

1. For protocols rather than TCP, like UDP, because they don't have built-in feedback loop, which means that all traffic can be one direction. In this scenario, we can't wait until there are packets available to piggyback the confirmation.
2. For TCP protocol, we don't simply populate a new TCP packet and send back because the sequence numbers of each direction are way different. Through our experiments, some NAT devices will refuse to transfer this packet if the sequence number messes up for the same TCP connection. To solve this problem, we add the request into a waiting list. When there are TCP packets available to send, we make a copy of the packet and attach the CM block with *CM Flags* of *Flags_Enabled* to send out. In this case, there will be two consecutive TCP packets with the same sequence number. The NAT devices will simply consider them as retransmissions instead of dropping them.

For both the MPIP query packet and MPIP confirmation packet, we populate new packets based on the original packets by making copies. All copied packets won't be passed to higher layers, they are generated at network layer and stop there too, they don't mean anything to higher layers. The network layer drops all MPIP packets with *CM Flags* of *Flags_Enable* and *Flags_Enabled* after processing them. The MPIP handshake process is shown in Figure 1.

With the handshake flow above, the smallest number of query packet that will be sent for the whole process is only one. But sometimes because of packet loss or synchronization issues, there can be multiple query packet sent out at both ends as shown in Figure 1. This is not a problem because the design of the system allows receiving more than one query packet and confirmation packet. But on the other hand, for nodes that don't support MPIP, we can't send out query messages forever. In Table 2, the column *Query Count* maintains the number of query messages that have been sent out to relative IP addresses. If the number is larger than a threshold value, it assumes that IP address doesn't support MPIP. Considering that packet loss is rare event in Internet nowadays, we set this value to 5 in our system, this is more than enough to make sure a MPIP enabled node can receive

and reply this message with successful transmissions. In a connection with only one-way traffic, if the receiver is MPIP enabled while the sender isn't, then no query packet will be sent out at all.

Generally, for MPIP enabled nodes, they may have more than one IP address. In this situation, multiple synchronization messages will be transmitted for the same node.

3.3 Path Management

Given M and N interfaces at each end of one connection, there are totally $M * N$ possible paths on the connection. In our system, we maintain all available paths through Table 3.

Node ID is used to identify the other end of connections. Path ID is used to identify one path. It is a locally unique Arab numbers starting from 1. This value is generated only when adding new paths into Table 3.

Because of NAT mapping, the same node can have different IP address and port number that will be seen on the other side at different time or even for different application. That's why column *Session ID* needs to be added into Table 3. When sending out one packet, the final path will only be chosen from the ones that have the correct session id. The detail of session is explained in Section 3.7. Here we just consider a session is the synonym of one connection. As the names show, source IP, source port, destination IP and destination port show the address information of the path.

One instance of Table 3 is maintained at each side. Initially, Table 3 is empty. Every time one new combination of IP address and port number is flagged as MPIP enabled in Table 2, it adds new paths to the table with every available local IP address as the source IP address, and the value of Node ID is extracted from Table 4 which is maintained to conveniently map node ID with IP address and port number.

Table 4: Node ID vs IP address and Port

Node ID	IP Address	Port Number
ID_1	IP_{11}	P_{11}
ID_1	IP_{12}	P_{12}
ID_2	IP_{21}	P_{21}
ID_2	IP_{22}	P_{22}

For multipath connections with full mesh, if both M and N are larger than 1, every NIC is reused at least once. Sometimes it is not necessary to reuse them because of system overhead consideration; In this use case, we can limit the number of available paths for one session to be l , then only the fastest l paths will be used, other paths will be abandoned.

3.4 Feedback Loop

Packet loss and delay are the most important two characters of one path. Given that packet loss in current Internet has become rare events, we only focus on delay in our system. Furthermore, different paths can be chosen on different side, round trip delay doesn't work for MPIP. Based on this,

we implement IP layer's feedback loop with one-way delay as the feedback variable.

In Table 3, all fields referring to network delay will be filled or calculated by feedbacked one-way delay. In Table 1, the fields *Path ID* and *Packet Timestamp* are used to measure network delay. When node A sends out a packet, it chooses a path from Table 3, fill the field *Path ID* with the chosen path ID, and fill the field *Packet Timestamp* with local system time T_1 in the CM block. After node B receives this packet, it extracts node ID, path ID and timestamp from the CM block. The node ID and path ID are directly used to identify records in Table 5, and node B uses $T_2 - T_1$ as the one-way delay from node A to node B where T_2 is the local system time when receiving the packet. Node B checks whether the path that identified by the node ID and path ID already exists in Table 5, if yes, it updates the path's delay with $T_2 - T_1$, otherwise, it adds a new record into Table 5.

Table 5: Path Feedback information

Node ID	Path ID	Path Delay	Feedback Time
ID_{11}	PID_{11}	D_{11}	T_{11}
ID_{12}	PID_{12}	D_{12}	T_{12}
ID_{21}	PID_{21}	D_{21}	T_{21}
ID_{22}	PID_{22}	D_{22}	T_{22}

In practice, the value of path delay calculated here isn't the real delay value because of time difference between node A and node B as shown in 4.1.1, it can even be negative. But as we will see later, time difference between A and B doesn't have any influence on our algorithm.

When node B needs to send packet back to node A , it chooses the record with the earliest feedback time from Table 5, fills the field *Feedback Path ID* and *Path Delay* in the CM block, and updates the column *Feedback Time* with local system time. When node A receives this feedback packet, it extracts the path ID and path delay value, and fills the path delay value into the column *Real-Time Network Delay* in Table 3. To avoid outliers, the value of path delay is calculated by moving average algorithm.

3.4.1 Dynamic Addition/Removal of Paths

As multipath feature enabled on a device, IP addresses of devices can be removed or added dynamically. Especially for mobile device, it can connect to different access points(WiFi hotspot/Cellular Tower) at different time, during this stage, its IP address can be changed, removed or added dynamically. Under this situation, the system supports dynamic addition and removal of paths from Table 3.

When IP address change happens, the value of *CM Flags* is set to be *Flags_IP_Change* in the CM block when sending out packets. After receiving packets of this flag, the receiver knows that IP address change happened on the sender, it will remove all entries of this connection in all tables, then add the combination of IP address and port number con-

Table 3: Path information

Node ID	Path ID	Session ID	Src IP	Src Port	Dest IP	Dest Port	Minimum Network Delay	Real-Time Network Delay	Real-Time Queuing Delay	Maximum Queuing Delay	Path Weight
ID	PID_{11}	SID_1	SIP_1	SP_1	DIP_1	DP_1	D_{min11}	D_{11}	Q_{11}	Q_{max11}	W_{11}
ID	PID_{12}	SID_1	SIP_1	SP_1	DIP_1	DP_2	D_{min12}	D_{12}	Q_{12}	Q_{max12}	W_{12}
ID	PID_{21}	SID_2	SIP_1	SP_2	DIP_1	DP_1	D_{min21}	D_{21}	Q_{21}	Q_{max21}	W_{21}
ID	PID_{22}	SID_2	SIP_1	SP_2	DIP_1	DP_2	D_{min22}	D_{22}	Q_{22}	Q_{max22}	W_{12}

tained in the received packet into all the tables as if that combination is MPIP enabled. Also, the sender does the same reset for this connection. After all these resets, there is only one path left for this connection, all the other paths will be added into the system through the normal process. By doing this path reconstruction, we achieve smooth transmission during IP address changes which is transparent to higher layers.

3.5 Periodical Heartbeat

For protocols like TCP, during the whole lifetime of the connection, both sides are sending packets to each other at a high frequency, then at both sides, Table 3 can be updated real-time. But there are protocols that don't have this built-in feedback mechanism, like UDP. In some UDP applications, all traffic is one way, there aren't any acknowledgements, which means that the sender can't get feedback information through piggybacked messages. Under this scenario, the sender won't be able to properly add new entries into Table 3, then multipath feature can't be applied at all.

To solve this problem, a periodical heartbeat mechanism is introduced into the design. At each side, when the node receives packets, the system checks Table 5 for the specific node, if it finds the feedback time of this path is close to a pre-set value, it makes a copy of the received packet, switches the source/destination address information, and sends back the packet with CM block attached. The path to send this heartbeat message will be chosen through the same algorithm as regular IP packet as in Section 3.6. Through the heartbeat message, we effectively maintain the real-time status of each path between two nodes for connections with one-way traffic. The expiration time of heartbeat message is set to 300ms in our system.

All heartbeat packets have a special flags value $Flags_HB$ in the CM block. These packets will be dropped after being processed at network layer.

3.6 Path Selection

Every time one node needs to send out packets, it chooses the most suitable path from Table 3. Depends on the requirement of throughput and responsiveness, we have different considerations to choose the target path.

3.6.1 Delay-based Path Selection

In the current Internet, there are many high delay-bandwidth product connections. These connections generally have both high delay and high bandwidth. For applications that want to

achieve high throughput, we propose following mechanism to achieve the requirement.

The criterion of choosing the best path that can achieve high throughput is on the column of *Path Weight*. Given certain values of this column, we don't simply choose the path that has the largest path weight which may overuse the path and starve other paths. Instead, we choose the path by random number. For one specific path k , the probability $P(k)$ it will be chosen is calculated in Equation 1. By balancing the percentage of packets on each path, system fluctuation can be effectively avoided.

$$P(k) = \frac{W_k}{\sum_{i=1}^N W_i} \quad (1)$$

From above, the path weight is the only criterion to choose the target path, so the calculation of path weight W is critical to the performance of the system. Wrong decision can result in catastrophic disaster. In our prototype, by introducing delay-based solution, we calculate the value of W in an incremental pattern.

For network delay, it consists of following 4 parts.

1. Processing delay. Time routers take to process the packet header
2. Queuing delay. Time the packet spends in routing queues
3. Transmission delay. Time it takes to push the packet onto the link
4. Propagation delay. Time for a signal to reach its destination

Among the four parts above, processing delay, transmission delay and propagation delay are fixed value for one path, they can be treated as constant. But queuing delay changes as the traffic load on the path changes. Figure 2 shows the trend of network delay as of the traffic on the path changes.

We can see that the queue on the path starts to accumulate at point C , but still, there is no loss. Starting from point L , the queue size overflows and packet loss happens. The minimum and maximum network delay for the path can be measured at point C and point L . The difference between the real-time delay and minimum delay can be treated as the approximate queuing delay which represents congestion situation on that path. By observing the real-time queuing delay, we can adjust the weight of each path to properly assign future packets.

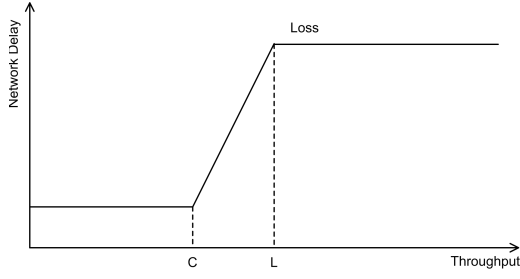


Figure 2: Network delay trend as throughput increases

In Table 3, except the column *Real-Time Network Delay*, other three delay related columns are calculated through real time delay D .

1. **Minimum Network Delay D_{min} .** Every time one node receives update of network delay, it update this column with the minimum of its current value and the new real-time network delay.
2. **Real-Time Queuing Delay Q .** According to Figure 2, the value of queuing delay is calculated as $Q = D - D_{min}$ which is the difference between real-time network delay and minimum network delay.
3. **Maximum Queuing Delay Q_{max} .** Maximum queuing delay is updated once the real-time queuing delay Q is larger than Q_{max} . When packet loss happens, this value notates the curve section after point L in Figure 2.

During our experiments, we found that calculating the weight of each path independently according to queuing delay can result in high fluctuation. So instead, we couple all the paths together and do micro adjustment to the weight of each path periodically as shown in Algorithm 1.

Algorithm 1 Path Weight Incremental Adjustment.

```

1:  $Q_{avg} = \frac{\sum_{i=1}^N Q_i}{N}$ ;
2: if  $Q_i \leq Q_{avg}$  then
3:    $W_i = W_i + S$ ;
4:   if  $W_i > 1000$  then
5:      $W_i = 1000$ ;
6:   end if
7: else
8:    $W_i = W_i - S$ ;
9:   if  $W_i < 1$  then
10:     $W_i = 1$ ;
11:   end if
12: end if
13: return;

```

In Algorithm 1, N is the number of paths that belongs to one specific connection, Q_i is the queuing delay of path i , W_i is the path weight of path i , and S is the adjustment

granularity. Initially, every path has the same path weight of $\frac{1000}{N}$. In each loop, the path weight increases or decreases by S . The maximum value of the weight is 1000. The minimum value is 1 because we try to keep all the paths alive in case the congestion on a bad path has huge relief.

Algorithm 1 is executed periodically, the length of of period is defined as a configurable system variable T . So here, we have two configurable parameters which are the granularity S and the period T . There is a trade-off here. Larger S and shorter T can generate higher fluctuation but faster convergence while smaller S and longer T can result in slow convergence and lower fluctuation. In our system, according to the path weight range (1 1000), we set S to 10 and T to 100 milliseconds. During our evaluation, this configuration can converge fast with low fluctuation.

3.6.2 Customized Path Selection Framework

In practice, maybe high throughput can represent most user requirements, but in some special cases, responsiveness can be of the first priority.

A typical example is Skype[17] calls. For Skype, audio packets and video packets are transmitted independently. In most scenarios, with responsive audio quality, real-time video streaming can be a bonus. But large delay in audio streaming can be a nightmare for a Skype call. In this case, obviously, audio packets and video packets have different priorities. During our study of Skype packets, although video packets can also be small, audio packets generally have substantially shorter length comparing to video packet. This provides a chance to define specific rules for Skype audio packets to achieve better user experience with lower delay.

Also, in a TCP connection with multipath, ACK packets are generally very small. On the other hand, delay in ACK packets can trigger TCP congestion control which results in unnecessary degradation of performance. By sending small ACK packets on the path with lower delay, unnecessary congestion control can be avoided, and further improve the overall performance of the connection.

To address the requirements above, we enable the users to define their own customized path selection policy based on the destination and length of packets. This customization provides a fundamental framework for more advanced path selection algorithms. A dedicated MPIP routing table is defined in Table 6.

Table 6: MPIP routing table

IP Address	Port Number	Protocol	Start Size	End Size	Routing Priority
*	23	TCP	0	200	P_{Res}
192.168.1.2	80	TCP	200	*	P_{Tp}
192.168.1.3	5221	UDP	0	500	P_{Res}

When sending a packet, the system checks the destination IP address, port number, protocol and length of the target

packet to get relative routing priority from Table 6. Different routing priority has different path selection policy. In our system, only two priorities are supported. *P_Throughput* means throughput is the first priority, the delay-based path selection algorithm in Section 3.6.1 will be used. When *P_Response* is referred, it means responsiveness is the first priority. In this scenario, the path that has the lowest delay will be chosen to send out the packet.

In the first row of Table 6, for any TCP connection with destination port 23, if the packet length is smaller than 200 bytes, the path with the lowest delay will be chosen, otherwise, delay-based algorithm is used. The second row actually is useless because all packets will use the delay-based algorithm even this row doesn't exist. The third row specifies that UDP packets sent to 192.168.1.3's 5001 port will be assigned to the lowest delay path, otherwise, delay-based algorithm is used.

The same as regular routing table, the content of the MPIP routing table is configurable by users. In MPTCP, because all congestion control algorithms inherit from traditional TCP, it can't make the best of multipath. On the other hand, MPIP maintains paths in a more customized pattern, it is feasible to have application specific routing decisions. As we mentioned above, we only provide a basic framework for customized routing in MPIP. It only has limited functionalities. For example, it only specifies applications by port number. For some applications like P2P software, they have arbitrary port numbers, we can't locate the specific application in this case. There can also be other scenarios that our prototype doesn't work. But based on this framework, a more powerful and smarter routing decision mechanism can be completed.

3.7 Session Management

For one specific connection, it can be identified by a socket pair at each end. Each socket pair is described by a unique 4-tuple consisting of source and destination IP addresses and port numbers. In our prototype, we call a connection's unique socket a session. To maintain all sessions, each MPIP enabled node maintains an instance of Table 7.

For each connection, the system internally maintains a socket pair with source/destination IP addresses and port numbers, the system maintained socket information is the same as the pair maintained in Table 7. The socket information for a session entry won't be modified after they have been added into Table 7 even the IP address that initiates the session doesn't exist any more. As will be shown later, the socket information in Table 7 is used to communicate with higher layer to guarantee seamless connection switching. If these information is modified, higher layers will notice the mismatch between the system maintained socket and MPIP maintained socket and causes connection failure.

3.7.1 Addition/Removal of Sessions

Session ID is the unique identity of one session. Unlike path ID in Table 3 which only requires to be unique locally,

session ID needs to be unique on both ends of one connections which means that the session ID needs to be the same on both ends for the same session. But from Table 7, we can see that different nodes can have the same session ID. Same as path ID, session ID is Arab numbers starting from 1. This value is generated only when adding new sessions into Table 7.

After the MPIP availability handshake has been successfully completed, when sending out a packet, the sender checks Table 7 to see whether a proper session entry has been generated. If not, one new record will be added. We extract the IP addresses, port numbers and protocol from packet headers, and gets the destination node ID from Table 4, then it generates a new session ID and adds one new entry into Table 7. After this, the new session ID will be filled into *Session ID* in the CM block whenever packets are sent out on this connection. The receiver extracts the session ID and inserts one entry into its own Table 7. This will make sure that for the same session, both sides of the connection have the same session id because only one side generates it.

Besides the session ID, all IP addresses and port numbers can be different in Tabel 7 of both sides because of NAT devices, this doesn't cause any problem because the IP addresses and port numbers only need to be the values that are seen locally.

Removal of sessions is decided by expiration. At each node, every time it sends or receives a non-heartbeat packet, it goes to Table 7 to get the session ID to fill the control message and updates the column *Update Time*. For an active session, this time stamp should be updated frequently. If the timestamp expires a threshold value, the session is considered to be obsolete and removed from Table 7. In our system, this threshold value is set to 120 seconds. After all information related to this session is removed, if there are still more packets coming on that connection, MPIP handshake will start over.

Once session entry has been inserted into Table 7 at both side, different paths will be added into Table 3 and available to be chosen to transmit future packets that belong to this session according to path selection algorithms.

3.7.2 NAT Consideration

NAT devices can be the largest obstacle for MPIP. There are many types of NAT devices. Different type of NAT device does different modifications to Internet traffic. Generally, a NAT device changes the IP addresses and port numbers of a packet. But further, some application-level gateways can modify more content of a packet like the sequence numbers of TCP packets. We can still name many other different cases where NAT devices do customization to packets. In these scenarios, MPIP won't be able to perform correctly. Instead of designing an all-purpose MPIP system, we choose to fall back to regular connection in these case.

The value of the column *Update Time* in Table 7 is also used to monitor the correctness of MPIP. As mentioned, this

Table 7: Session Information

Node ID	Session ID	Source IP	Source Port	Destination IP	Destination Port	Protocol	Next Sequence No	Update Time
ID_1	SID_1	SIP_1	$SPORT_1$	DIP_1	$DPORT_1$	TCP	S_1	T_1
ID_1	SID_2	SIP_1	$SPORT_2$	DIP_1	$DPORT_2$	UDP	0	T_2
ID_2	SID_1	SIP_2	$SPORT_3$	DIP_2	$DPORT_3$	TCP	S_2	T_3
ID_2	SID_2	SIP_2	$SPORT_4$	DIP_2	$DPORT_4$	UDP	0	T_4

value is updated when packets are received on this session. In our system, if this value is not updated for 120 seconds, this session is considered to be dead. But if this value is less than 120 seconds and larger than 500ms, we assume that MPIP is not working correctly, then this session falls back to regular connection and multipath feature is disabled. We set the value of 500ms which is slightly larger than the period of heartbeat.

This mechanism works perfectly for connections that transmit packets with high intensity, e.g., chunky file download, video/audio streaming, etc. If for some reason the packets don't reach the destination at all through MPIP, the timer expires soon and falls back to regular connection. Specifically for TCP traffic, sometimes the packets reach the destination with modified header which the receiver can't recognize, the sending window of the connection becomes zero, traffic on both direction pauses soon, and finally the session falls back to regular connection.

But for applications that transmit packets intermittently, the connection may fall back to regular connection falsely. For example, a SSH session may idle for a long time because of no traffic at all. For these scenarios, we don't do any special process in our system, we just fall back as designed because if a connection is not active enough, the benefit it can get from MPIP is trivial.

Through this simple algorithm, we monitor MPIP's performance as a blackbox. MPIP can fail because of many factors on NAT issue, but no matter why MPIP is acting incorrectly, regular connection is always there to take over.

3.7.3 Workflow of Sending/Receiving Packets

For one regular packet, when being sent, it goes through application layer, transportation layer, then arrives at network layer where IP protocol resides in. Before the packet arrives at network layer, socket information inside the packet is the one that maintained by OS. In regular scenarios, the system finds a proper interface to push out the packet according to the destination IP address and routing table. The opposite flow is executed for inbound packets. For a MPIP enabled system, all the other operations remain the same except that of network layer.

For outbound packets, when the packet arrives at network layer from transportation layer, the system looks into Table 7, and finds the session entry that matches socket information in the original packet, extracts the session ID to fill into field *Session ID* in CM block. To choose the proper path

to send out the packet, the first thing is to locate which paths in Table 3 are eligible for this connection. Given the destination IP address and port number in the socket information, we can find *Node ID* in Table 4, then in Table 3, all entries that have the correct node ID and session ID are eligible. Among all these paths, the most suitable path will be chosen out through the mechanism introduced in Section 3.3.

The chosen path's path ID will be assigned to field *Path ID* in CM block for delay measurement at the receiver, meantime, we modify the source and destination IP address/port number in the IP header and transportation layer header with the chosen path's source and destination IP address/port number. Then we route the packet with the new IP header information to a proper interface.

For inbound packets, when receiving a packet, the receiver extracts node ID and session ID in the CM block, with these two parameters, the receiver locates the original socket pair in Table 7, and modify network layer header and transportation layer headers with source and destination IP address/port number from Table 7. Fields related to path measurement will be processed as explained in Section 3.3. Now the packet is back to its original shape that can be recognized by higher layers, it is ready to be pushed up.

3.8 Multipath TCP Transmission

Through our experiments and previous studies[1], NAT devices have a lot of interferences to end-to-end connections, especially for TCP packets. The most straightforward limitation for TCP packets is that many NAT devices will drop TCP packets that don't have a connection related to them. MPTCP doesn't have this problem because each subflow in MPTCP is a regular TCP connection. But in MPIP, if we transmit TCP packets on a path rather than the original one, NAT devices on the path will probably drop these packets before they arrive at the destination. To solve this problem, in our MPIP implementation, we provide two options to solve this problem.

3.8.1 Fake TCP connection

In NAT devices that drop TCP packets without relative connection information, we cheat them by constructing fake TCP connections. The construction of TCP contains three-way handshakes. Instead of constructing real TCP connections, we implement a simple three-way handshake at network layer which is similar as TCP handshake, but this connection information won't go up to TCP layer. All hand-

shake packets have *CM Flags* value of *Flags_Hs*, these packets are dropped after being processed by MPIP.

As shown in Table 1, the field *Local Address List* carries all local IP addresses. Also, the node that initiates the connection is considered as the client. When the client receives the IP address list of the server, it extracts its own IP address list, then sends out a SYN packet through each possible path to the server except the original one which is the one that was used to initiate the connection. When the server receives this SYN packet, it replies with a SYN-ACK packet through the same path. After the client sends out the final ACK packet to the server, the three-way handshake for our fake TCP connection is completed successfully. After this, the path can be used to transmit TCP packets without being dropped.

In one specific TCP connection, we assume that the server has at least one public IP address for initiating the connection. But it is possible that other IP addresses on the server are not public, then the SYN packet will never arrive at that address, in this case, this interface will not be used at all.

3.8.2 UDP wrapper

The other option we provide for multipath TCP traffic is UDP wrapper. Because there is no connection information in an UDP packet, during our experiments, most NAT devices don't have any limitation on regular UDP traffic. We make use of this feature to wrap our TCP packet with a UDP header to pass the NAT devices and unwrap it when received.

At the sender side, every time the network layer gets a TCP packet from transport layer, the system chooses a path to send the packet out as shown in Section 3.3. If the chosen path isn't the original path, we wrap the user data and TCP header into an UDP packet.

If UDP wrapper is used to transmit TCP traffic, when receiving one packet, we are capable to know this UDP packet is a wrapper for a TCP packet instead of a regular UDP packet by checking the column *Protocol* in Table 7. After removing the UDP wrapper, socket information will be extracted from Table 7 and filled into the TCP and IP header.

3.8.3 Dealing with out of order TCP traffic

In modern devices, different interfaces can have totally different delay behaviour. Generally, cellular interface has much larger delay than WiFi. For PC devices, wired connection generally has smaller delay than WiFi interface. Under this situation, packets can be out of order frequently during the life time of a connection. This is not a problem for protocols like UDP, but for TCP, as shown below, packet out-of-order can result in degradation of overall performance.

In TCP Congestion Avoidance algorithm, a retransmission timer expiring or the reception of duplicate ACKs can implicitly signal the sender that a network congestion situation is occurring. The sender immediately sets its transmission window to one half of the current size. When a duplicate ACK is received, the sender doesn't know if it is because a TCP packet was lost or simply that a packet

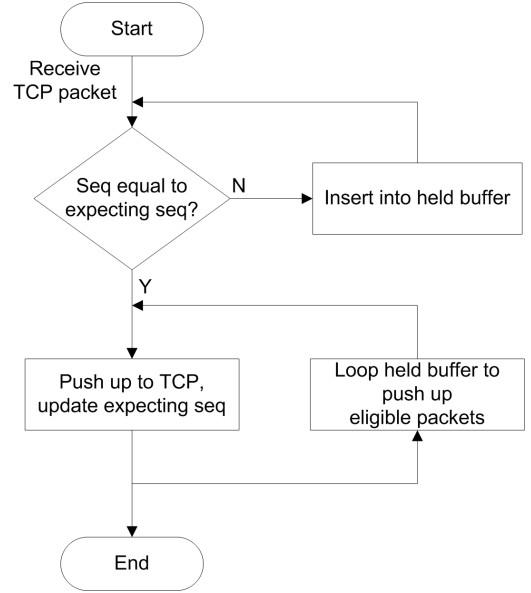


Figure 3: TCP Out-Of-Order Packet Process

was delayed and received out of order at the receiver. If the receiver can re-order TCP packet, it shouldn't be long before the receiver sends the latest expected acknowledgement. Typically no more than one or two duplicate ACKs should be received when simple out of order conditions exist. If however more than two duplicate ACKs are received by the sender, it is a strong indication that at least one packet has been lost, the sender does not even wait for a retransmission timer to expire before retransmitting the segment and enter control avoidance stage, the sender's transmission window is cut off by one-half. This is the fast retransmit algorithm.

Considering multipath scenario, if the delay behaviour difference among different paths is not trivial, we can expect a lot of out-of-order packets. This will result in many fast retransmit events, even TCP can go back to slow start stage. Given that in modern Internet, real packet loss has become rare event, most of transmission window cut-off events are unnecessary at all. In our prototype, to solve the heterogeneous delay performance of different paths, we make specific process of TCP out-of-order packets.

The overall process of out-of-order packet is shown in Figure 4(b).

For each session in Table 7, if it is TCP protocol, MPIP maintains a buffer *B* to store all out-of-order packets and the next expecting sequence number *S*. Every time one node receives a TCP packet, it takes out the sequence number, if it equals with the current expecting sequence number of that session, the packet is pushed up to transportation layer immediately, also, the buffered packets will be checked to see whether there are qualified packet for delivery. If not, the packet will be buffered into *B* for future delivery. Every time one packet is pushed up to transportation layer, the next expecting sequence number *S* is updated through Equation 2

where L is size of the whole packet, H_{ip} is IP header size, and H_{tcp} is TCP header size. TCP sequence number ranges from 0 to $2^{32} - 1$. Equation 2 can overflow S . But the data type of sequence number is unsigned integral, the number will loop back to 0 automatically after that.

$$S(k) = S(k - 1) + (L - H_{ip} - H_{tcp}) \quad (2)$$

For the data structure that holds the TCP packets, a proper choice is to implement it as a binary tree, but according to our observations, most out-of-order packets come in an incremental order, and they are held in the buffer because of one late packet, when that late packet arrives, all the held packets will be pushed up. For this reason, we simply implement this structure a sorted list, every out-of-order packet will be inserted into this list in ascending order according to the sequence number. The time complexity to insert this packet is almost $O(1)$ because most out-of-order packets arrive in an incremental order as mentioned. When the expecting packet arrives, then all the waiting packets can be cleared from this list with a simple loop. This reduces code complexity greatly without sacrificing performance.

During our experiments, it happens that one specific packet can be late for a long time, like if the packet is lost. In this situation, holding all subsequent packets will halt the whole session because TCP layer will assume that all packets are lost, this will result in catastrophe for the connection. To address this problem, we set up the maximum size of the buffer. All the packets in the buffer will be forcefully pushed up once it is full. In our prototype, we set this maximum size to 10.

3.8.4 MPIP and MPTCP work together

As the first implementation of multipath, MPTCP gains huge attractions in research. During the development of our prototype, we try to do side-by-side comparison between MPTCP and MPIP when referring to TCP connection. Besides this, for investigation purpose, we merge them together to see how the system works.

Assume that there are 2 NICs at each end, MPTCP will have 4 subflows for the session while each subflow is an independent TCP connection. MPIP will have four paths for all these four subflows which means that there will be totally 16 paths for one session in this case.

When MPIP and MPTCP combine together, if packet loss happens, MPIP will notice the high queuing delay of the dedicated path, and avoid assigning packets to that path by reducing the path weight to prevent the situation becoming worse, then subflows of MPTCP triggers TCP congestion control less than regular case. Adding with MPTCP's unique two-level sequence number mechanism, the load balancing among paths at MPIP enables MPTCP to make better use of each path and generate higher throughput. We will see how this combination works in Section 4.

4. PERFORMANCE EVALUATION

To evaluate the performance of our proposed system, we implemented our multipath IP in Linux kernel 3.12.1 under Ubuntu system. We evaluate the performance of the system in both controlled lab environment and Internet. Also, we verify our customized routing framework with UDP traffic in controlled lab environment.

In all experiments, we try to keep the configuration of each node unchanged after installation. We don't do any special configuration to the system, neither we do any optimization to squeeze out all possible throughput. Except specific experiments that can only be applied to MPIP like UDP experiment and customization MPIP routing, we try to do side-by-side comparison with MPTCP for TCP connections. We will figure out how these two features work independently and together as stated in Section 3.8.4;

For all throughput related experiments, we use iperf3 to generate traffic between the client to the server.

4.1 Controlled Lab Environment

In the lab, we install the prototype on two desktops. Both desktops are connected directly to one router without any middle-box. Each desktop has two 100Mbps NICs which means that there are totally 4 paths and the capacity is 200Mbps between the two nodes. We use Netem tool in Linux to throttle the connection to evaluate our prototype under multiple scenarios. Wireless connection is also considered in our evaluation.

4.1.1 Clock Offset

During experiments, we found that the clock of each node has some small difference. In our configuration, the clock of the server is slightly faster than the client. Even this error is very small, we still see the difference in a long experiment. Given most computers have NTP enabled and the system's local time synchronize with time server periodically, but we still think that this difference is worth to be shown here.

We turn off NTP on both nodes, do a TCP transmission for one day with consistent traffic and record the one-way delay of each packet from the client to server. Because the traffic load is consistent, queuing delay roughly remains the same. But as shown in Figure 4(a), because of the clock offset between the two nodes, the trend of queuing delay exposes an linearly increasing curve even the trend is very slow. In Figure 4(a), we record the queuing delay every one minute. For the whole day(1400 minutes), we can see that the clock offset is about 350 milliseconds which means that the server's clock runs one millisecond fast than the client for each four minutes. We will be able to see this trend again in following results.

4.1.2 TCP Out-of-order Process

In Section 4.1.2, we explained why out-of-order is a problem that must be dealt with in multipath implementations, and we also proposed our solution to this problem. To ver-

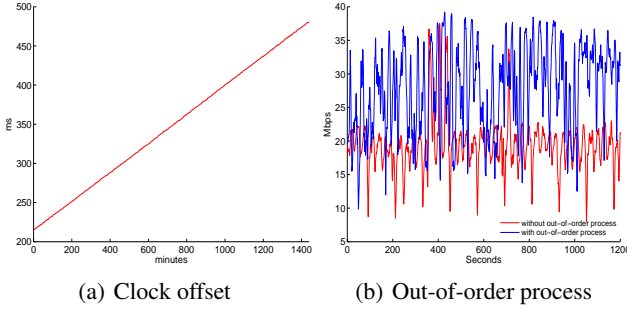


Figure 4: Clock and Out-of-order

ify our proposition, we replace one NIC card on the client with a wireless interface. In our experiment plat, the RTT of one path with two wired NICs is about 0.1ms while paths with one the wireless NIC is about 0.5ms which will generate enough out-of-order packets. Also, to make sure that there are heavy load of packets to be assigned to the wireless NIC card, instead of using the standard path selection algorithm in Section 3.6, we fix the weight of all the four paths to be the same. Then we make sure that 50% of outbound packets will be assigned to the wireless NIC. With this configuration, we do a regular TCP transmission that lasts for 20 minutes with out-of-order process enabled and disabled respectively. The result is shown in Figure 4(b).

With the same configuration, Figure 4(b) shows the improvement brought by the out-of-order process. The average throughput is 28.2Mbps and 19.4Mbps with/out out-of-order process respectively. The improvement maybe trivial if the delay on all the paths is the same because most packets will arrive at the receiver in the order of being sent out. But for multipath connections, it can be very often that each path goes through a totally different route, that is where out-of-order happens most. In all following experiments, we enable out-of-order process by default.

4.1.3 Dynamic Addition/Removal of Path

In Figure 5, a 6 minutes experiment is done for MPIP and MPTCP respectively. The experiment consists of three stages. We start the experiment with 4 paths fully functional. After two minutes, we unplug the wire of one NIC at the client to reduce the number of path to 2. The wire is plugged back for the last two minutes to add back the number of path to 4.

In both case, MPIP and MPTCP adapt to the change smoothly. The transmission of iperf3 doesn't pause besides the change of throughput. Because of the overhead to close and open new TCP connections, when IP address changes happen, MPTCP has longer delay than MPIP, especially when adding paths.

4.1.4 TCP Throughput Enhancement

As we mentioned in Section 3.8, there are two different implementation of multipath TCP in our system to solve

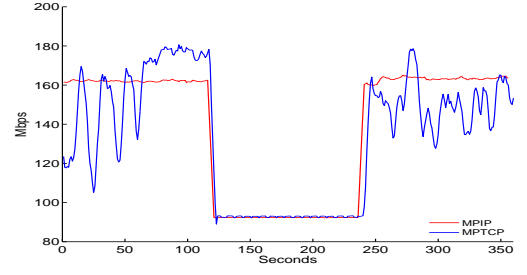


Figure 5: Dynamic Addition/Removal of Path

NAT problem which are fake TCP connection and UDP wrapper. We do a specific experiment to evaluate the performance of each approach. With a 20 minutes TCP transmission for each, the average throughput for fake TCP and UDP wrapper is 165.4Mbps and 164.7Mbps respectively while we get an average throughput of 92.7Mbps for regular TCP. This result shows that both implementations have roughly the same performance. In all following experiments, we use fake TCP by default, but UDP is still kept as an option for users.

Now we start to do side-by-side comparison between MPIP and MPTCP for TCP traffic. Figure 6 shows the average throughput comparison results for multiple configurations.

We first do the transmission without any throttles which means the capacity of the connection is 200Mbps. We can see that MPIP achieves the highest throughput which is 171Mbps, MPTCP only gets 129.5Mbps. When we combine MPIP and MPTCP together as stated in Section 3.8.4, we get a throughput of 164.6Mbps.

By limiting the bandwidth from one NIC of the client to the two NICs on the server to 20Mbps, we create a connection with capacity of 140Mbps. MPIP gets the lowest bandwidth here followed by MPTCP, but when they work together, the highest throughput is achieved.

In our experiment plat, with wired connection, the round trip time is trivial(about 0.1ms). To emulate a connection with more delay, we manually add 5ms delay to each NIC on the client and get result of the fourth bar group in Figure 6. The result is roughly the same as without limit.

By replacing one NIC on the client to a wireless interface, we evaluate the performance of MPIP under wireless connection. In this case, MPTCP achieves higher throughput than MPIP, but still, when they work together, the highest throughput is achieved.

The error bar on top of each bar is the standard deviation the related bar. To avoid unnecessary errors, we neglect all data of the first 100 seconds to wait the traffic of each path stabilizes. The overall variation of one session is smaller than independent paths' variation. This happens because the coupling among paths causes variations for one single path, but with every path compensating for each other, the overall variation is small. The two wireless paths have small absolute variation value, but comparing to their small throughput, the value of variation is actually very large.

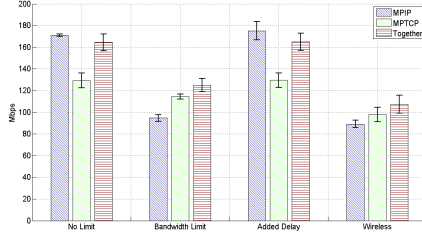


Figure 6: Overall Throughput Comparison

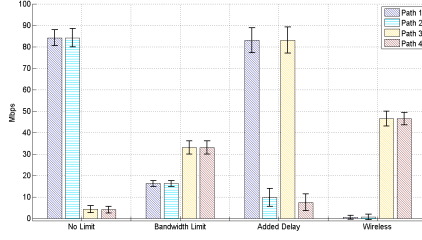


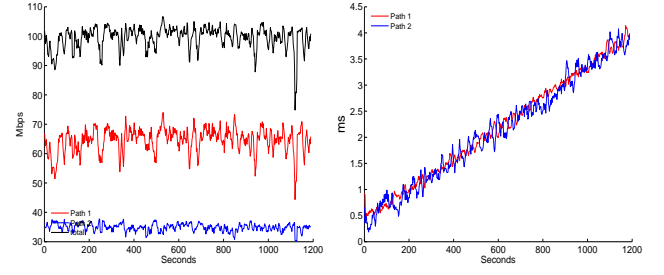
Figure 7: Path Throughput Comparison

In Figure 7, we do a analysis among different paths for MPIP transmission to clarify how all packets are assigned to each path.

From the throughput comparison results, we can see that even the characters of a path is almost the same, because the coupling of different paths, there can be huge difference in their throughput. In our experiment set-up, in the case of no limit and adding delay, all four paths literally have the same characters, but results show that most traffic is taken by two paths. This happens because the four paths are not independent while one NIC is shared by two paths. The queuing delay generated by one path may also affect the queuing delay on the other path. When limiting bandwidth, we can see that the two limited paths(1 and 2) don't use all their bandwidth, neither the unlimited two paths(3 and 4), this causes the low throughput for MPIP in this scenario shown in Figure 6. When wireless is applied to the connection, the two paths that contains the wireless interface almost get nothing assigned. This is because the high and unstable delay of these two paths.

In Figure 8, we limit the number of path to 2. The two paths don't share any NIC to prevent coupling among paths. Also, we limit the bandwidth of path 1 to 80Mbps and path 2 to 40Mbps. From the throughput trend in Figure 8(a), both paths converge to their own configured bandwidth and stay stable for the whole experiment. Figure 8(b) shows the queuing delay information trend, the influence of clock offset is still straightforward. According to calculation, the amount of offset during the whole experiment is consistent with the result shown in Figure 4(a). Generally the queuing delay keeps stable once the TCP connection passes slow-start.

For paths that have wireless interface, the fluctuation is



(a) Throughput Comparison (b) Queuing Delay Comparison without Limit

Figure 8: Path Throughput and Path Queuing Delay

Table 8: Wireless MPIP Queuing Delay

	Path 1 (Wireless)	Path 2 (Wireless)	Path 3 (Wired)	Path 4 (Wired)
Average (ms)	4.15	3.84	1.88	2.09
Standard Deviation	13.03	11.27	0.79	0.80

very large. Besides their average value is almost two times of wired connection, their variation is even larger as shown in Table 8. In Figure 7, we can see that the two wireless paths get little traffic because the large queuing delay values generate small value of path weight.

4.2 On the Internet and 4G

Besides the controlled lab experiments, we also evaluate MPIP on the Internet to verify the prototype's NAT immunization and system robustness. We set up our server at Emulab[18] which is located in Utah while the client is in New York. Because there is only one NIC that connects to Internet on the Emulab node, there are only 2 paths in this case.

With this configuration, we do three experiments for different bottleneck placement. First, we connect both NICs to the same router that connects to Optimum Cable with 15Mbps bandwidth. In this case, both NICs share the same bottleneck, we don't expect any throughput enhancement at all. Secondly, we connect one NIC to Optimum and connect the other NIC to Verizon FIOS with 25Mbps bandwidth. Finally, we connect one NIC to Optimum and connect the other wireless NIC to TMobile 4G network through an iPhone 6 plus's hotspot. In the latter two cases, each NIC has different bottleneck, and potential throughput enhance is expected. According to our test, the capacity between our client and the server in Emulab is about 5Mbps for Optimum, 10Mbps for FIOS and 900Kbps for TMobile 4G. The result of iperf3 TCP transmission result is shown in Figure 9 and Figure 9.

When sharing bottleneck, all three sessions fully use the capacity of the connection with little difference. The two paths don't share the traffic equally either because of coupling between them. When Optimum and FIOS are used,

overall, MPTCP gets the lowest throughput among the three sessions followed by MPIP. Still, when MPIP and MPTCP combine, the highest throughput is achieved. As for the assignment of packets for the two paths in MPIP, Path 1 in Figure 9 is the path for Optimum and path 2 is for FIOS. Approximately the traffic is assigned to each path according to the capacity of each link. This proves that the delay-based path selection algorithm in Section 3.6.1 also works perfectly in real Internet.

The result for 4G connection is slightly different. In this case, MPTCP gets the highest throughput, the other two sessions get roughly this same throughput which is lower than MPTCP. According to our observation, comparing to Optimum and FIOS, the delay of 4G connection is very unstable, this causes very high queuing delay on the 4G path. According to the delay-based path selection algorithm, for most of time, the weight of the 4G path is very small. This can be shown in the traffic assignment for the 4G path in Figure 10. When MPIP and MPTCP work together, both TCP sessions have two paths at IP level, neither makes good use of the 4G path because of the unstable delay.

When both paths share the same bottleneck, the two paths are assigned with different amount of traffic because of coupling between the two paths, that's why we see the same difference of variation as in Figure 6 and 7. Both overall variation and single path's variation get large value when the two paths have different bottlenecks.

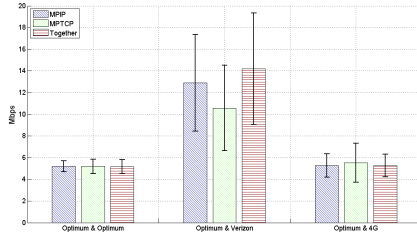


Figure 9: Overall Throughput Comparison in Internet

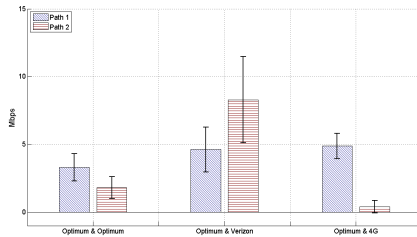


Figure 10: Path Throughput Comparison in Internet

4.3 UDP and Customization Routing

In this section we try to evaluate the usage of UDP and customization routing in our MPIP system. We won't include any throughput enhancement here because the result

is straightforward that we get nearly double throughput by applying multipath for UDP traffic in our blab experiment plat. Instead, we use Skype[17] as an experiment application to show how UDP works in our system. Meantime, we will show how our customization routing mechanism helps improve Skype audio call quality.

Skype uses direct connection for two parties calls. All the video and audio packets are transmitted between two ends of the call directly through UDP protocol. Comparing to video streaming, Skype calls generally have higher requirement on the responsiveness of audio packets. According to our experiments, almost all Skype audio packets are less than 200 bytes while video packets are generally larger than 1000 bytes. To make optimization to audio packets, we add one entry to Table 6 to assign packets smaller than 200 bytes responsiveness priority. Besides a Skype audio call, we also run an iperf3 TCP session between the two nodes to see how TCP and UDP coexists in MPIP.

For this experiment, we only use one NIC on the server while we still have two NICs working on the client. For path 1, we set the bandwidth to 2Mbps and the delay to 50ms. For path 2, we set the bandwidth to 300Kbps and the delay to 20ms. Then we have one high bandwidth-delay product path and one low bandwidth-delay product path. By enabling and disabling customization routing, we get the result in Figure 11. Figure 11(a) shows the throughput of iperf3 TCP connection with/out customization, Figure 11(b) shows Skype's audio round trip time during the whole experiment with/out customization. This round trip time information is extracted from Skype's own real-time technical report, it shows the amount of delay the user experiences during the audio call.

From Figure 11(b), we see the huge reduction of audio delay. The average RTT is 82ms and 119 ms with/out routing optimization. With queuing delay based algorithm, most Skype audio packet will be assigned to path 1 because of its high bandwidth, but the by-product is higher delay. By assigning audio packets to path 2, a much better audio call quality can be achieved. Meantime, from Figure 11(a), we see that we get roughly the same TCP throughput in both case, but the result with customization routing enabled is more consistent. This is because with customization routing enabled, path 2 has its designated traffic, the queuing delay accumulates because of its low bandwidth, then TCP traffic will probably be assigned to path 1 and results in a more consistent throughput.

In Figure 11(c), by applying customization routing, we prove that this mechanism can also improve TCP throughput. TCP ACK packets are generally very small. If these packets are assigned to a high delay path, probably, TCP congestion control will be trigger and pull down the overall performance of TCP.

We make the configuration same as the one for Figure 4(b) by replacing one wired NIC with wireless NIC and assigning all the path weight to the same value. In Figure 11(c), we do

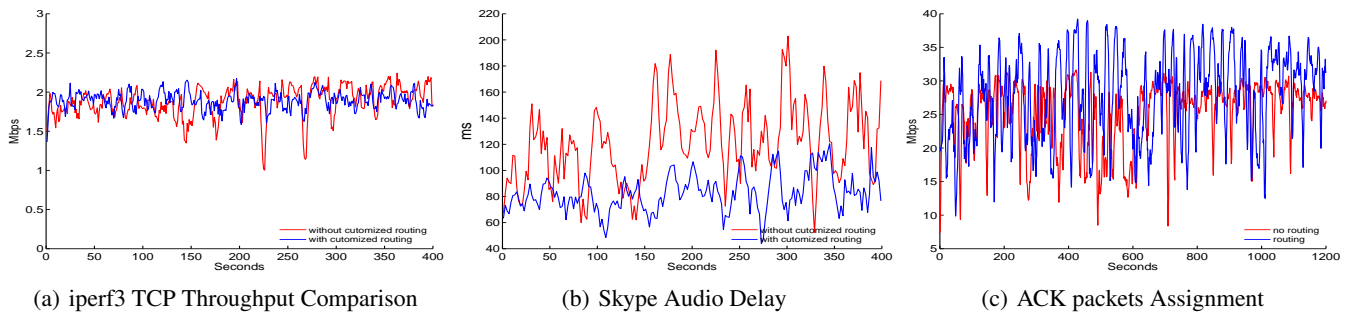


Figure 11: UDP and Customized Routing with Skype and ACK Optimization

the same TCP transmission by enabling and disabling customization routing for small packet as in Skype. We can see obvious throughput enhancement in Figure 11(c). By calculation, the average throughput of the two cases are 24.2Mbps and 28.5Mbps respectively.

5. CONCLUSIONS AND FUTURE WORK

Given the popular support of multiple network interfaces on modern devices, MPIP is designed to make full use of each interface to achieve optimal performance. Besides higher throughput than single path transmission, customized routing for specific application's requirement is also in the consideration of MPIP when designing the system. Instead of sole support of TCP traffic, MPIP works with any upper protocol that uses regular IP as the transmission method, i.e. TCP and UDP.

In the work, we prove that multipath can be implemented at network layer with less effort and better maintainability because of the simplicity of network layer. According to our evaluation, besides higher throughput than regular TCP, customized routing capability can greatly benefit applications that have different requirement of high throughput and high responsiveness. By combining MPTCP and MPIP together, with the support of MPTCP's unique two-layer sequence number management and independent TCP subflow mechanism, we surprisingly achieved the best throughput result.

We can see that with unstable connections like 4G, MPIP's performance is yet satisfactory. We use a simple delay-based path selection algorithm in this prototype, a more complete and efficient module can be designed for this functionality in the future. Also, customized routing now only supports packet length based decision, a smarter decision mechanism like flow-table in SDN is under consideration. As future works, Android implementation is next step to directly use 4G interface instead of hotspot access. IPv6 support is also another feature that is underway.

6. REFERENCES

- [1] C. Raiciu, C. Paasch, S. Barre, A. Ford, M. Honda, F. Duchene, O. Bonaventure, and M. Handley, "How hard can it be? designing and implementing a deployable multipath tcp," in *NSDI*, 2012.
- [2] H.-Y. Hsieh and R. Sivakumar, "A transport layer approach for achieving aggregate bandwidths on multi-homed mobile hosts," in *MobiCom*, 2002.
- [3] P. Key, L. Massoulié, and D. Towsley, "Path selection and multipath congestion control," *Commun. ACM*, vol. 54, no. 1, Jan. 2011.
- [4] Y. Dong, D. Wang, N. Pissinou, and J. Wang, "Multi-path load balancing in transport layer," in *Next Generation Internet Networks, 3rd EuroNGI Conference on*, May 2007.
- [5] O. L. C. Cien, and Y. J., "An introduction to the stream control transmission protocol (sctp)," RFC 3286, 2002.
- [6] I. Joe and S. Yan, "Sctp throughput improvement with best load sharing based on multihoming," in *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*, Aug 2009.
- [7] S. Barre, C. Raiciu, O. Bonaventure, and M. Handley, "Experimenting with multipath tcp," in *SIGCOMM 2010 Demo*, September 2010.
- [8] C. Paasch, G. Detal, F. Duchene, C. Raiciu, and O. Bonaventure, "Exploring mobile/wifi handover with multipath tcp," in *Cellnet*, 2012.
- [9] G. Detal, C. Paasch, S. van der Linden, P. Merindol, G. Avoine, and O. Bonaventure, "Revisiting flow-based load balancing: Stateless path selection in data center networks," *Computer Networks*, vol. 57, no. 5, April 2013.
- [10] C. Paasch, R. Khalili, and O. Bonaventure, "On the benefits of applying experimental design to improve multipath tcp," in *CoNEXT*, 2013.
- [11] C. Paasch, S. Ferlin, O. Alay, and O. Bonaventure, "Experimental evaluation of multipath tcp schedulers," in *ACM SIGCOMM Capacity Sharing Workshop (CSWS)*, 2014.
- [12] Y.-C. Chen, Y.-s. Lim, R. J. Gibbens, E. M. Nahum, R. Khalili, and D. Towsley, "A measurement-based study of multipath tcp performance over wireless networks," in *IMC*, 2013.
- [13] Y. Cao, M. Xu, and X. Fu, "Delay-based congestion control for multipath tcp," in *ICNP*, 2012.
- [14] L. S. Brakmo, S. W. O'Malley, and L. L. Peterson,

“Tcp vegas: New techniques for congestion detection and avoidance,” in *SIGCOMM*, 1994.

- [15] D. Thaler and C. Hopps, “Multipath issues in unicast and multicast next-hop selection,” RFC 2991, 2000.
- [16] Apple, “Apple,” <http://www.apple.com/>.
- [17] Skype, “Skype,” <http://www.skype.com/>.
- [18] Emulab-Team, “Emulab - Network Emulation Testbed Home,” <http://www.emulab.net/>.