

TCP Performance Enhancing Proxy Support For Non-interactive Applications

Joe Bayer

Informatics: Programming and System Architecture
60 ECTS study points

Department of Informatics
Faculty of Mathematics and Natural Sciences

Joe Bayer

TCP Performance Enhancing
Proxy Support For
Non-interactive Applications

Supervisors:
Michael Welzl
Kristjon Ciko

Contents

1	Intro	2
1.1	Motivation	2
1.2	Research Question	3
1.3	Contributions	3
1.4	Organization	3
2	Background	4
2.1	TCP/IP	4
2.1.1	Congestion Control	5
2.1.2	3 Way handshake (0 RTT)	6
2.1.3	TCP Options and Fast Open	7
2.2	The future of wireless communication.	8
2.2.1	5G Millimetre Wave	8
2.2.2	Buffering	9
2.2.3	Non-Interactive Applications	11
2.3	Proxy	11
2.3.1	PEP	12
2.3.2	PEP for wireless communication	12
2.3.3	Transparent vs Non-Transparent	12
2.4	Linux	13
2.4.1	Kernel Modules	14
2.4.2	C Programming Language	14
2.5	Related Work	15
3	Design	17
3.1	Justification for designing a PEP	17
3.2	Performance	18
3.2.1	Programming Language	19
3.2.2	Kernel Module Vs. Userspace Application	19
3.2.3	Connection splitting using Sockets	19
3.2.4	Bandwidth utilization	20

3.2.5	PEP Selection	22
3.2.6	Connection Establishment	22
3.3	Deployment (Transparency)	24
3.3.1	Transparent PEP	24
3.3.2	Non Transparent PEP	25
3.4	Buffering	26
3.4.1	Socket buffers	26
3.5	AQM & Scheduling	27
3.6	Security	27
3.7	Summary	27
4	Implementation	29
4.1	Kernel Module	29
4.1.1	Kernel Hooks	30
4.1.2	Linux Version and Distribution	31
4.2	TLV Library	32
4.2.1	Custom connect function	33
4.2.2	TLV	34
4.2.3	TLV Options	35
4.2.4	Shared Library	36
4.3	PEP - Internals	36
4.3.1	Architecture	36
4.3.2	Kernel Sockets	38
4.3.3	Work Queues	40
4.3.4	Works	41
4.3.5	Kernel TCP receive and send	44
4.4	PEP - Server	44
4.4.1	Creation	44
4.4.2	Server initialization	45
4.4.3	Accept and Endpoint connection	47
4.4.4	Multiple Servers	49
4.5	PEP - Clients	49
4.5.1	Client Sockets - Endpoint Sockets	49
4.5.2	PEP Connections	51
4.5.3	Module Customization	51
4.5.4	System Configurations	52
4.5.5	Userspace	53
4.6	Memory	53

5	Evaluation	54
5.1	Traffic Control Options	54
5.2	Ocarina Testbed	54
5.2.1	Hosts and Hardware	55
5.3	Initial test	56
5.3.1	Configuration	57
5.3.2	Experiment Procedure	59
5.3.3	Results & Analysis	60
5.4	Using real 5G traces	64
5.4.1	Traces	64
5.4.2	Topology	65
5.4.3	Configuration	65
5.4.4	Validation	65
5.4.5	Experiment Configuration	67
5.4.6	Experiment Steps	69
5.4.7	Results	71
5.4.8	Evaluation of 5G traces experiment	80
5.5	Socket buffer size effect on throughput	82
5.6	Multiple flows	83
6	Conclusion & Future Work	86
6.1	Performance	86
6.1.1	Throughput	86
6.1.2	Impact on interactive traffic	87
6.1.3	Transparency	87
6.2	Limitations and Challenges	88
6.3	Future work	88
6.4	Final Thoughts	89

List of Figures

2.1	Example of network domains	5
2.2	The TCP handshake procedure	7
2.3	Capacity fluctuations from mmMave from real 5G traces [25]	9
2.4	Linux packet queuing	11
2.5	PEP installed to support Wireless traffic over satellite.	13
3.1	PEP within a network (Simplified)	17
3.2	Example 5G network topology	18
3.3	Example of poor bandwidth utilization.	20
3.4	Example of good bandwidth utilization.	21
3.5	Split connection with bandwidth and delay.	21
3.6	The TCP handshake procedure across PEP	23
3.7	Optimal handshake across PEP (0 RTT)	24
3.8	Deployment of a PEP at a base station	26
4.1	Example of TLV header and entries	35
4.2	The architecture of the PEP	37
4.3	PEP State Structure Code	37
4.4	Kernel Sock Structure	39
4.5	Socket Protocol Operations from Linux - net.h [18]	40
4.6	Work struct from Linux - workqueue.h	41
4.7	https://docs.kernel.org/core-api/workqueue.html	47
4.8	Work operation table	51
4.9	Callback function table	52
5.1	Overview of the Ocarina testbed	55
5.2	Screenfetch results	56
5.3	FQ-CoDel traffic control visualization	58
5.4	BFIFO traffic control visualization	59
5.5	File transfer impact on interactive traffic	60
5.6	Close up of End to End compared to PEP	62

5.7	Sender to Router and Router to Receiver effect on the interactive traffic.	63
5.8	mmWave capacity fluctuation for four different 5G scenarios (Data from [25])	65
5.9	BFIFO traffic control visualization for 5G traces	66
5.10	Results from the Witestlab experiment [25]	66
5.11	Our results for the delay caused by fluctuating bandwidth	67
5.12	Testbed configuration for 5G traces	68
5.13	Calculation of the Bandwidth-Delay Product (BDP) in Bytes	69
5.14	Cubic's congestion window	70
5.15	Reno's congestion window	71
5.16	Bar plot of throughput for each scenario with a static link configuration	72
5.17	Impact on interactive traffic for the static link scenario	73
5.18	Closeup of the PEP and FQ-CoDel impact on interactive traffic for the static link scenario	74
5.19	Bandwidth fluctuations with Short blockage scenario	74
5.20	Bar plot of throughput for each scenario with a short blockage configuration	75
5.21	BFIFO impact on interactive traffic with short blockage and static link.	76
5.22	Bar plot of throughput for each scenario with a long blockage configuration	77
5.23	Bandwidth fluctuations with mobility and blockage scenario	78
5.24	Bar plot of throughput for each scenario with the mobility and blockage scenario	79
5.25	Impact on interactive traffic for each configuration in the mobility and blockage scenario.	80
5.26	Closeup of the PEP and FQ-CoDel impact on interactive traffic for the mobility and blockage scenario	81
5.27	BFIFO impact on latency.	81
5.28	Average throughput for each scenario by configuration.	82
5.29	Throughput for the PEP based on socket buffer sizes in ten steps.	83
5.30	The throughput of ten parallel flows for each configuration.	84
5.31	Impact on interactive traffic for ten flows with mobility and blockage scenario.	85

List of Listings

2.4.1 Listing 2.4.1: Default C program.	15
4.1.1 Listing 4.1.1: The basic kernel module setup code.	30
4.1.2 Listing 4.1.2: Example of a TCP congestion controller module	31
4.3.1 Listing 4.3.1: Work initialization example	41
4.3.2 Listing 4.3.2: Work using containerof example	42
4.3.3 Listing 4.3.3: Accept callback function	43
4.3.4 Listing 4.3.4: Forwarding callback function	44
4.4.1 Listing 4.4.1: PEP server creation	45
4.4.2 Listing 4.4.2: PEP server initialization (Simplified)	46
4.4.3 Listing 4.4.3: PEP server accept function (Simplified)	48
4.5.1 Listing 4.5.1: Client Forwarding Function (Simplified)	50

Abstract

This thesis proposes a novel solution to optimize network performance in fluctuating bandwidth environments, with a particular focus on 5G mmWave technology. The primary objective is to develop a Performance Enhancing Proxy (PEP) optimized for non-interactive traffic, aiming to maximize bandwidth utilization without disrupting interactive traffic. Using large buffers within the network, our approach uses socket buffers and active queue management to improve performance. Through an evaluation with real life 5G traces, we demonstrate the effectiveness of the PEP both in performance and impact on interactive traffic. Ethical considerations regarding the deployment of transparent PEPs are also addressed, ensuring user consent and privacy are upheld.

Acknowledgments

I would like to acknowledge and thank my supervisor Micheal Welzl for motivating me throughout my thesis and being always available for feedback and discussions. And my co-supervisor Kristjon Ciko for helping me navigate the enormous Linux kernel and restart the testbed machines when I inevitably would crash them.

I would also like to thank my friends, co-workers and family for keeping my spirits high. Additionally, I would like to thank my work place for being understanding and giving me the time and resources for my thesis. At last I want to thank my wonderful girlfriend for motivating me and helping me proofread hundreds of pages.

Chapter 1

Intro

1.1 Motivation

At an Internet bottleneck, the capacity that is available to a transport protocol can fluctuate for various reasons - traffic fluctuations, user mobility, or new wireless technology where the underlying capacity changes. 5G technology has come and introduced increasingly large bandwidth capacities[3]. However, because of the mmWave technology, the bandwidth experiences drastic fluctuations from any kind of line of sight blockage[23].

These fluctuations result in underutilization of available bandwidth during peak phases, significantly affecting throughput and overall network performance. Addressing this issue requires an innovative approach to large buffers and split connections, ensuring high performance without degrading the interactive traffic.

Current network strategies struggle with the optimal use of fluctuating bandwidth nature of 5G mmWave technology[25]. Specifically, packet schedulers and buffer configurations struggle to fully utilize the high bandwidth periods, leading to considerable performance losses. The challenge is to devise a Performance Enhancing Proxy (PEP) that can implement larger buffers and split connections effectively. Importantly, this solution must overcome the pitfalls of bufferbloat[11], which particularly impacts interactive traffic, and address the ethical concerns associated with the use of transparent PEPs that modify connections without user consent.

1.2 Research Question

When the capacity at a bottleneck fluctuates, senders can react as soon as a signal reaches them, i.e. in the order of a round-trip time; in the meantime, packets are lost. When capacity suddenly becomes available, however, it can take much longer for a sender to be able to transmit enough data to the bottleneck and capacity may only be available for a short period of time. It is therefore necessary to buffer data near the bottleneck. However, large buffers have been known to create latency problems (commonly known as "bufferbloat").

Accordingly, the main research question is:

Can we get better performance without a negative influence on latency by adding large buffers inside highly fluctuating networks?

1.3 Contributions

In this master thesis we aim to better understand and utilize larger buffers in the network and present evidence that large buffers do not inherently cause bufferbloat or negatively impact interactive traffic. Additionally, because of the challenges with extending TCP[12], this thesis investigates the field of Performance Enhancing Proxies with focus on transparency. The aim is to remove some of the stigma associated with Performance Enhancing Proxies which are used today.

1.4 Organization

The thesis is organized in the following manner: Chapter 2 introduces required background knowledge for the thesis such as 5G mmWave, the Bufferbloat concept and Performance Enhancing Proxies. Chapter 3 will explore the design decisions made regarding the programming language, environment, platform and deployment. Chapter 4 details the implementation and configuration.

In Chapter 5 we will look at the evaluation of the Performance Enhancing Proxy. The evaluation will consist of multiple experiments to confirm the effectiveness of the Performance Enhancing Proxy itself. The final chapter, Chapter 6, will conclude the thesis based on the evaluation and look at potential future work.

Chapter 2

Background

In this chapter we will present some of the required background knowledge to understand the concepts presented in this paper. Focusing on topics that are outside the common understanding of network programming, especially details of certain congestion controllers and network protocols will be discussed. The rest of the thesis will assume the following topics are known to the reader.

2.1 TCP/IP

Perhaps the most well known internet transport protocol is the Transmission Control Protocol (TCP). It is known for providing reliable and in-order delivery of packets using acknowledgments and re-transmissions [9]. It was first introduced in 1974, but is still one of the most used internet protocols. However, as the demands of the internet have changed, TCP has not. Though TCP has been updated with minor extensions over the years, such as an increased initial window or new options, the core ideas have stayed the same [4].

TCP often spans multiple different domains with varying topologies and demands, especially between wired and wireless domains:

- **Wireless Domain:** A wireless communication domain refers to the transmission of data over a wireless medium without the use of physical connections such as wires or cables between devices. This domain covers a variety of technologies, including 3G, 4G, and 5G for mobile communication, Bluetooth and Wi-Fi for close-range communication, and satellite communication for worldwide communication.
- **Wired Domain:** Unlike a wireless domain, a wired domain provides a

steady and reliable bandwidth with low error rates and high throughput. The use of Ethernet and Fiber are typical for wired networks, they enable the transmission of a large amount of data over long distances with low signal noise.

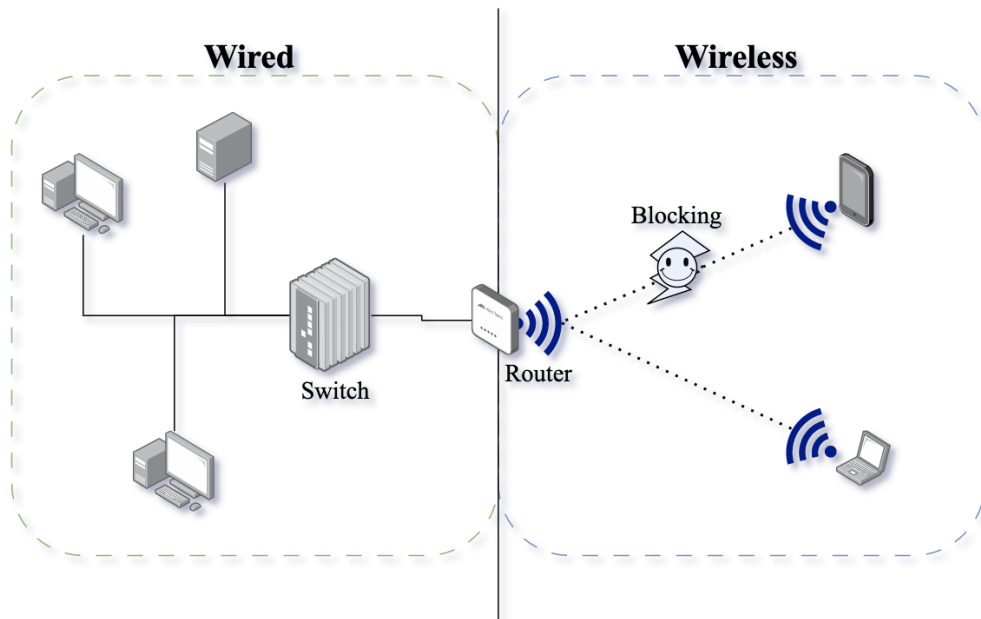


Figure 2.1: Example of network domains

Each domain has different requirements that a single TCP connection cannot satisfy. Fig. 2.1 shows the two domains and their characteristic differences. Usually, wireless domains experience a lot of changes in connectivity and bandwidths, while the wired domain is usually considered stable. This creates problems for TCP which normally spans multiple domains. Especially congestion control has problems adapting to highly fluctuating bandwidth across long distances and multiple domains.

2.1.1 Congestion Control

Congestion occurs in the internet when a network's resources, such as routers, are overloaded to the point that they diminish quality of the network [29]. Packet loss and delays are common issues associated to congestion in the network. To solve the problem of congestion, a distributed algorithm is used: Congestion Control. The main goal of congestion control is to maintain a stable network, while still utilizing the available bandwidth shared among all flows. This is achieved by for example: additively increasing the sending

rate, and multiplicatively reducing the sending rate when detecting congestion [28]. Congestion can be detected by monitoring packet loss, changes in delay, but also by explicit notifications.

Over time different variations of congestion controllers have emerged. Although their goal is the same: reduce congestion in the network, their approaches vary. Here are three examples:

- **TCP Reno:** Reno embodies the traditional approach to congestion control: slowly increasing the sending rate while the network is stable and drastically reducing it on packet loss. TCP Reno was designed for unstable and dynamic networks, where the rapid response rate is crucial to prevent network overloading. However, the slow start rate and aggressive reduction of the sending rate make it sub optimal for high bandwidth networks, as TCP Reno is not able to recover fast enough to utilize the high bandwidth after packet loss.
- **Vegas:** Vegas is similar to TCP Reno in most aspects, the main difference is the use of delay to detect congestion instead of packet loss. This makes New Vegas able to react faster to congestion, however it also introduces some problems. If Vegas competes with TCP Reno flows, it will start reducing its sender rate before TCP Reno does, this leads to Vegas losing out on possible bandwidth.
- **Cubic:** Cubic improves on the idea of TCP Reno by using a cubic function to adjust its sending rate in order to achieve higher throughput in a fast manner. Cubic is very efficient in highspeed networks and known for handling large data transfer over long distances. However, Cubic is not as reliable and robust as more traditional congestion controllers like TCP Reno, as the congestion window changes more rapidly.

In summary, the main differences between TCP Reno, Vegas and Cubic are their approach to congestion control, their performance in different types of networks, and their trade-off between efficiency and reliability.

2.1.2 3 Way handshake (0 RTT)

For TCP to establish a connection it uses a three-way handshake. Initially, it transmits a synchronization (SYN) packet to the desired endpoint. The endpoint responds with an acknowledgement and a synchronization packet of its own (SYN/ACK). Finally, the client responds with a acknowledgment (ACK). At this point both endpoints have confirmed that they are ready for

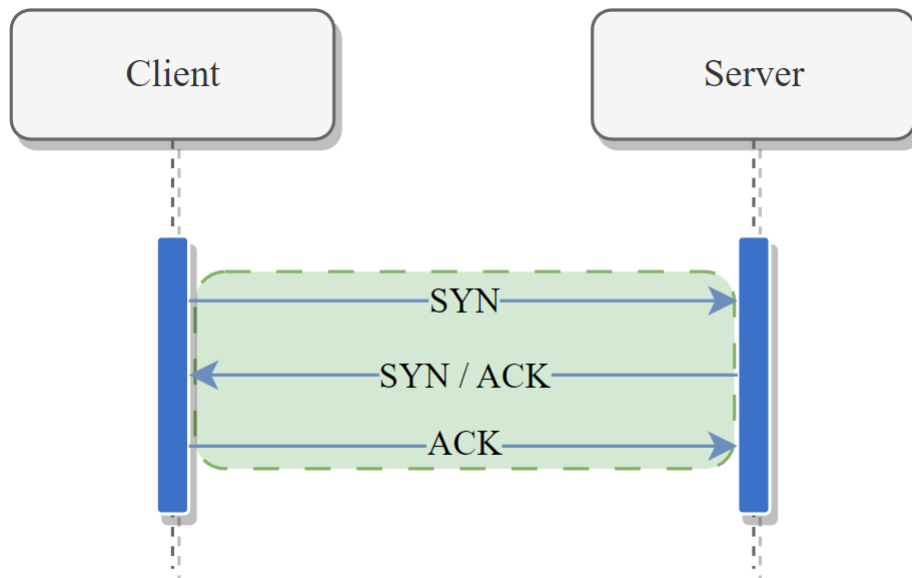


Figure 2.2: The TCP handshake procedure

further communication. For any connection to be established this handshake has to be done. For short flows that terminate in just a few round trips the initial TCP handshake can be a bottleneck, which is made worse if the connection is using a proxy and has to exchange additional information. The concept of proxies will be explored later in this chapter.

2.1.3 TCP Options and Fast Open

A TCP connection can be configured with optional header extensions called TCP Options [5]. These options change the default behaviour of TCP or add new features. One such feature is TCP Fast Open, which allows data to be added to the initial synchronization packet. A typical use case could be adding a HTTP GET request, thereby saving an entire round trip. In general, flows that terminate in a few round trips greatly benefit from this feature because the bottleneck is within the initial TCP handshake. Therefore, by removing the extra round trip required to send the first data packet, a significant amount of time can be saved.

TCP Fast Open also has other benefits, such as establishing connections to proxies [4]. When one is establishing a connection through a proxy, one gets the added delay of a second round trip for sending the desired endpoint. This can be avoided by using TCP Fast Open to send the desired endpoint

in the first synchronization packet to the proxy. "SYN forwarding" enables the users to establish a proxy connection without any added delays, however it does depend on the user's application to use TCP Fast Open.

2.2 The future of wireless communication.

Wireless communication has seen a lot of improvements such as highly increased bandwidth achieved through advanced technologies like 5G. Millimetre frequency bands have opened up new possibilities for wireless communication. These higher frequency bands offer greater capacity and can accommodate more devices, however high frequencies come with a set of new challenges such as highly fluctuating bandwidths. This fluctuation can be influenced by various factors such as signal interference, obstacles in the signal path, and environmental conditions [14].

2.2.1 5G Millimetre Wave

The emergence of 5G Millimeter wave communications has opened the doors for low latency networks with multiple gigabits bandwidth. This is achieved by using higher millimetre wave (mmWave) frequencies in the range of 30GHz to 300GHz, which has a lot of benefits [3]. A wider spectrum of frequencies to choose from and higher data transfer rates are just some of the many benefits mmWave provides. But alongside the benefits, mmWave has also introduced a lot of new challenges.

A big problem with millimetre wave communication is signal path blocking, also called "Line of sight blocking" [23]. It's caused by the use of Beam-forming to increase the bandwidth and range of millimeter wave signals. Beam-forming focuses the signal in a certain direction, making any blocking of the signal path devastating for the bandwidth. Even the human body can create enough blockage to drastically reduce the bandwidth. This causes immense fluctuations in the bandwidth whenever the signal is blocked.

Fluctuating bandwidths lead to unstable TCP connections with a worst case of losing packets. Current TCP congestion controllers such as CUBIC, Reno or Vegas struggle when reacting to sudden changes.[16] They are simply not able to utilize the high bandwidth when it is available. Increasing the aggressiveness of a congestion controller is not an option either, as it would disrupt the internet. A possible solution could be to buffer packets at the 5G base stations, having the data ready for when the bandwidth is high. This

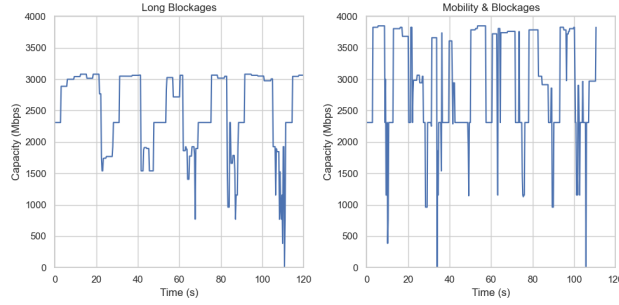


Figure 2.3: Capacity fluctuations from mmMave from real 5G traces [25]

however creates a new problem: "bufferbloat".

2.2.2 Buffering

Bufferbloat

The bufferbloat problem occurs when the systems between the endpoints buffer so many packets that the latency drastically increases and the reliability of the network as a whole goes down. The increased latency is detrimental for interactive (latency sensitive) applications. Generally it's preferred to drop packets and keep buffers small to avoid buffering time sensitive packets such as synchronization packets. Although this works in most cases, it's far from an optimal solution.

The increased bandwidth and low latency promises of new technology such as 5G has put a lot of pressure on the efficient forwarding of packets. Small buffers are therefore the standard, but at the same time, fluctuating bandwidth has shown the potential need to buffer packets for non-interactive traffic. Most focus has been on accommodating latency sensitive applications like virtual reality or remote surgery to name a few.

This thesis will explore non-interactive applications where latency is not that critical and more buffering is acceptable and most likely desirable. By splitting traffic into interactive and non-interactive we can improve the performance of both. By having very small buffers for interactive applications we avoid bufferbloat problems, while utilizing the benefits of big buffers for non-interactive applications.

Active Queue Management

Active Queue Management (AQM) is about managing the length of queues in a network. By dynamically adjusting the queue length, AQM algorithms[6] prevent the buffer from becoming too full or too empty[24]. This proactive management ensures smoother traffic flow and reduces the chances of packet loss or delay. AQM serves as a concept for more advanced techniques like packet scheduling, which further refines the process of handling network traffic.

Packet Scheduling

Another method of reducing the effects of bufferbloat is packet scheduling. A system should not send more packets than the weakest link can handle; this idea is built into TCP in the form of congestion control. However, when buffers grow to the point of causing bufferbloat, TCP's congestion control algorithms are unable to confidently determine a sending rate. Packet scheduling can solve this problem as it also controls the size of the buffers. It makes sure queues can grow when needed, but keeps the overall state of the buffers low. Packet scheduling has a lot more to offer than simple queue management, this will be explored later.

Proposed algorithms:

- **FQ-CoDel**: The Flow Queue Controlled Delay algorithm, FQ-CoDel for short, was developed to deal with the bufferbloat problem. Its main goal is to reduce the impact of head-of-line blocking and give a fair share of bandwidth by mixing packets from multiple flows [13]. Internally FQ-CoDel uses a FIFO queue, classifying packets into different flows to provide a fair share of bandwidth.
- **HTB**: Hierarchical Token Bucket is a queuing discipline based on assigning different classes a certain amount of bandwidth and sending rate. Because of its extensive bandwidth and delay management it's a good option for testing, especially in a virtual environment.
- **BFIFO**: Byte First-In-First-Out (BFIFO) is a simple queuing mechanism where packets are processed on arrival, without differentiating between traffic types. This approach can lead to increased latency for time-sensitive packets, especially under heavy, non-interactive traffic.

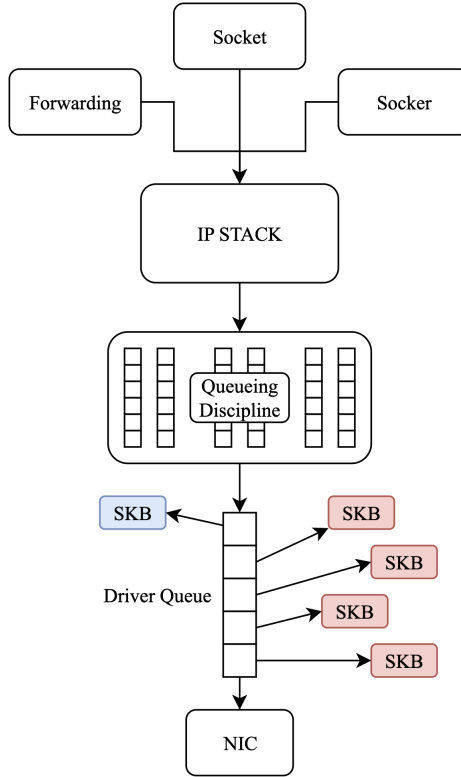


Figure 2.4: Linux packet queuing

2.2.3 Non-Interactive Applications

Non-Interactive applications such as web traffic, file transfers and video streaming can benefit from larger buffering, especially with fluctuating bandwidths. This is because, if we are able to buffer the packets closer to their final destination, we have them ready to be sent when the bandwidth changes. As this thesis will show, by utilizing larger buffers and AQM we can avoid unwanted delays and achieve better bandwidth utilization for non-interactive traffic. At the same time, with the solution presented in this thesis, interactive applications will not suffer under large queue delays that occur under large buffers.

2.3 Proxy

Proxy servers play a big role in the modern internet, delivering benefits such as anonymity and increased performance [27]. A common use case for a proxy is caching by keeping a copy of popular resources such as a websites. This

reduces the latency of accessing the resource as long as the proxy is closer to the user than the original copy. Locality plays an important role in the total latency as any transmission will always be limited by the speed of light.

A proxy can also be used for privacy similar to a Virtual Private Network (VPN). By redirecting network traffic through a proxy, the origin of the traffic appears to be the proxy server rather than the actual end-user. Hypertext Transfer Protocol (HTTP), a popular internet protocol used for accessing websites, has this functionality built in using HTTP tunnels and a special CONNECT method in its header:

```
CONNECT mn.uio.no/:22 HTTP/1.1
Proxy-Authorization: Basic encoded-credentials
```

2.3.1 PEP

A performance enhancing proxy (PEP) is a proxy designed to increase the performance of applications using it, typically by influencing the behavior of TCP. The idea behind the PEP is putting more logic, such as connection management, buffering, caching inside the network. As the name suggests, a PEP is designed to enhance the performance, but can also introduce new features to a network. An example of a new feature is the multipath support the TCP Transport converter gives [4].

2.3.2 PEP for wireless communication

Performance enhancing proxies are already deployed and in use for a lot of wireless communication, especially satellites and radio access networks [20]. They have an inherent performance increase just by splitting the connection between the wireless and wired domains. These PEPs are therefore often installed at the base stations. However they are unable to distinguish between interactive or non-interactive traffic, meaning their buffers need to be small to avoid bufferbloat and hence they still suffer from fluctuating bandwidth problems. Alternatively, a PEP with a large buffer can compensate for fluctuating bandwidth, but will introduce delay for interactive traffic.

2.3.3 Transparent vs Non-Transparent

A big discussion regarding PEPs has been if they should be transparent or non-transparent. Transparent PEPs are not visible to the applications that

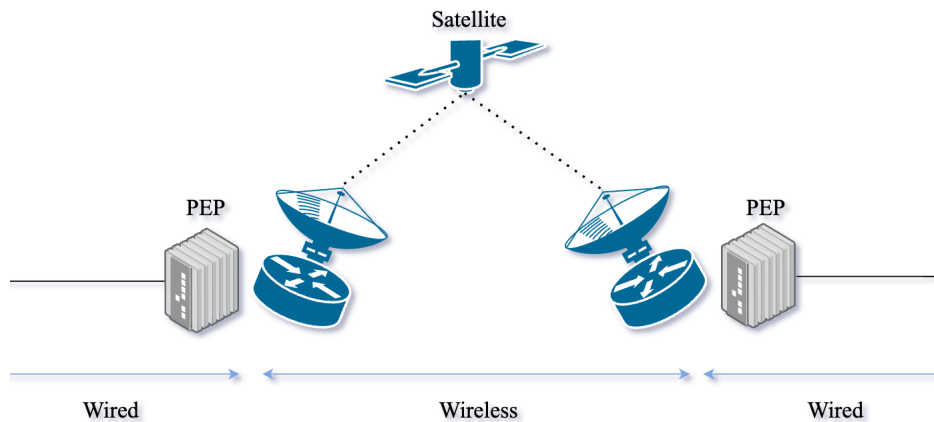


Figure 2.5: PEP installed to support Wireless traffic over satellite.

use it. They silently split the connections and spoof the IP-address of both the client and server [7]. This is prone to cause unintended side effects, such as certain TCP options not being forwarded, and security concerns. Non-Transparent PEPs on the other hand are explicitly chosen by either the client or the server, and the sender is aware of the proxy splitting the original connection. This approach can be seen as more appropriate for the internets architecture and could potentially remove some of the stigma associated with PEPs. This however requires modifications at the sender side to utilize the PEP.

2.4 Linux

Linux is the most famous open source kernel freely available for anyone to use and modify. Because of the open source nature of Linux, there have been many operating system implementations based on the Linux kernel. Ubuntu, Fedora or Manjaro are just some of the most famous Linux based operating systems out there. For developers, Linux is the perfect platform to experiment and test their new innovations. One is able to modify and recompile the kernel itself on the fly, and then test the solution on a live operating system. Linux supports most standards and is used by most major corporations such as Facebook, Amazon, Netflix and Google.

2.4.1 Kernel Modules

A concept that makes Linux truly extensible are Loadable Kernel Modules (LKM). Kernel modules are programs that can be loaded at runtime into the kernel and run with kernel privileges. Running with kernel privileges has a lot of benefits such as having access to internal structures and kernel symbols. Most drivers in the Linux kernel are written as kernel modules as they need access to the system internals.

Congestion controllers and packet schedulers are also usually implemented as kernel modules. That is because Linux exposes a struct with function pointers that can be overwritten by a module, making the kernel call the new functions instead. Because kernel modules run as part of the kernel they do not need to use a system call to do basic I/O as using sockets. Removing the overhead of system calls makes the kernel modules run much faster than default user space programs.

However, using Linux kernel modules has the drawback that the program is bound to Linux. The modules will only work in the context of the Linux kernel as they depend on the internal functions, and that they are part of the kernel. Most other operating systems like MacOS will not allow user defined modules to run with kernel privileges. Additionally, any bugs or error in the kernel module will make the entire kernel fail ("panic"), which usually requires a complete system restart to fix.

2.4.2 C Programming Language

C has been the optimal language for high performance systems since its creation in 1972 [15]. It was originally created for UNIX when it needed a higher level language, and now is the main programming language behind most operating systems such as Linux, Mac and Windows. Being very close to its predecessor, assembly, and compiled to a binary, makes it one of the fastest languages we have to date. Heap memory management is explicitly done by the programmer with no support for garbage collection. The unsafe memory management is one of the main challenges when programming in C, which however can be a benefit because a runtime garbage collection usually results in performance loss.

Listing 2.4.1: Default C program.

```
1 #include <stdio.h>
2
3 int main(void)
4 {
5     printf("Hello World!");
6     return 0;
7 }
```

2.5 Related Work

The 0-RTT TCP Convert Protocol is a protocol which aims to improve the performance for the fluctuating bandwidth and volatile nature of 5G networks by utilizing multipath TCP [4]. This is in contrast to our approach, which is to increase buffer sizes for the non-interactive traffic, although Multipath TCP is only one out of many possible use cases of the 0-RTT TCP Convert Protocol design. The protocol is non-transparent and uses TLVs to communicate with the proxy.

Other protocols, such as SOCKSv5, have bottlenecks when communicating with the proxy, because of the additional information that needs to be sent, such as authentication. The 0-RTT Transport Convert Protocol aims to reduce this overhead by utilizing "0-RTT", adding any meta-data to the SYN packet.

SnoopTCP is a similar proxy to our PEP, as it usually acts on the base station and aims to improve the control loops between the sender and the receiver. Different from our design SnoopTCP does not directly split the connection, but instead keeps a soft state of each TCP connection and re-transmits based on local timers. Additionally it suppresses duplicate ACKs on their way from the receiver to the sender [10].

SnoopTCP is transparent, which means it will split all connections and the sender will not be aware of any modifications SnoopTCP performs on the connection. The main benefit from SnoopTCP is its ability to react faster to packet loss with the local timers. However, because the congestion control is still end-to-end, the sender will not be able to send more data on demand if the bandwidth fluctuates.

Another similar proxy to our PEP is MilliProxy, which also focuses on 5G mmWave networks and aims to boost their performances. MilliProxy operates with a split connection, and keeps track of each flows state, having dynamic buffers per flow [21]. Similar to SnooPTCP, MilliProxy is transparent, which means it splits the connection without the knowledge of the sender or receiver.

PEP-DNA is a kernel module based PEP designed to aid deployments of new network architectures [7]. Similar to both SnooPTCP and MilliProxy, PEP-DNA is transparent. The most interesting feature of PEP-DNA for our PEP is the kernel module architecture, especially the split connection based on sockets and work handling. PEP-DNA uses a Netfilter to intercept SYN packets and then split the connection. This is a major difference to our approach which focuses on a non-transparent connection establishment.

Additionally, PEP-DNA does not distinguish between interactive or non-interactive traffic as all traffic is intercepted and the connections split. This essentially means PEP-DNA would not be able to perform the same optimizations with buffers and scheduling as our PEP can.

The Sidekick protocol is a new approach to PEPs which leaves underlying protocols opaque and unmodified [30]. Using a mathematical technique, named quACK, the Sidekick protocol is able to refer to packets without the access to cleartext sequence numbers. This approach comes as a response to new transport layer encrypted protocols such as Google’s QUIC. Similar to our design, the Sidekick protocol could be used to implement a split connection approach to increase performance of transfers which span multiple network domains. While our scheme is specifically designed for TCP, the Sidekick is generic; it offers an alternative way to implement the large-buffer scheme presented in this thesis, also supporting protocols such as QUIC.

Chapter 3

Design

In this chapter we discuss the design of a non-transparent connection splitting PEP at the boundary of an unstable network with fluctuating bandwidth, typically at 5G base-stations. By splitting the connection at this point we can buffer packets, usually closer to the user, and have them ready for high bandwidth phases. This is similar to how a content delivery network (CDN) stores caches and replicas closer to the users to reduce delay. The overarching goal is to improve the completion times of non-interactive traffic while avoiding to disturb the interactive flows that are passing through. A good design for the PEP is crucial as it needs to be robust, fast and reliable.

Figure 3.1 is a simplification of the PEP within a network, highlighting its interaction with other components, such as the client and endpoint. Importantly there is a fluctuating bandwidth between the client and the PEP.

3.1 Justification for designing a PEP

The ossification of networks, particularly TCP, has been a long-standing issue [12]. Over the years, the internet has evolved, but the core protocols,

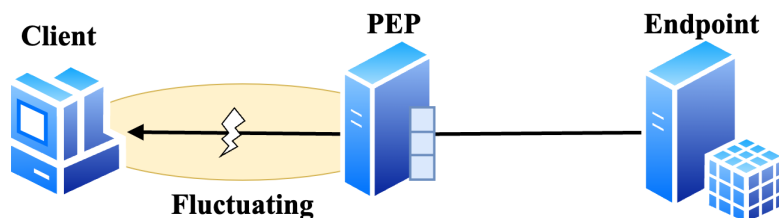


Figure 3.1: PEP within a network (Simplified)

like TCP, have remained relatively unchanged. This leads to challenges when attempting to introduce extensions or modifications. Altering such a fundamental protocol could disrupt countless systems and applications. This leaves us to explore new ideas using middle-boxes. A PEP is such a middle-box, as it can enhance the performance without introducing changes to TCP itself.

Using a PEP in combination with 5G has additional benefits. TCP with 5G needs to cover both the stable network and the fluctuating wireless domain split by the base station illustrated by figure 3.2. However, with a PEP we are able to split the domains and perform optimizations such as congestion control and buffering, tailored to each specific domain. Achieving the same optimization by modifying TCP would need a change to the tight integration of end to end congestion control.

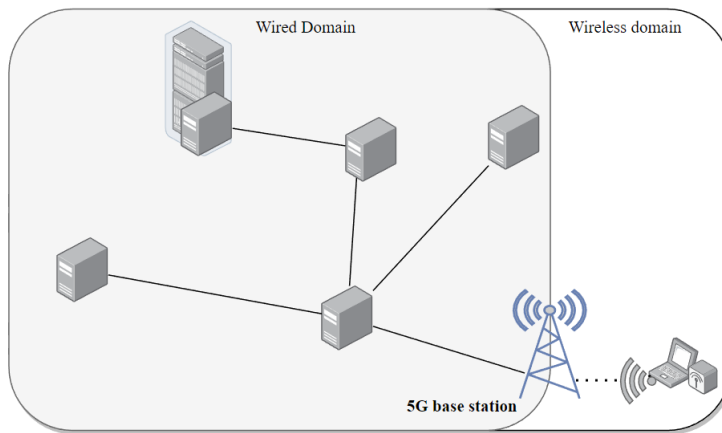


Figure 3.2: Example 5G network topology

3.2 Performance

In this section, we discuss the core design decisions to optimize the PEPs performance. The speed at which a PEP can both process and forward packets is crucial. Especially when wanting to utilize the rapidly fluctuating bandwidth of 5G, we need to react as fast as possible. The design choices presented in this section focus on the performance of the PEP. Because of these high performance requirements, the PEP is written in C as a Linux kernel module.

3.2.1 Programming Language

The programming language chosen for a PEP has a direct impact on its efficiency. Interpreted languages like Python might not offer the speed necessary for high-performance tasks. Even Java, while running within the JVM with JIT and garbage collection, can potentially introduce delays. The languages best suited for high performance are C, C++ and Rust. Both C and C++ are very similar and well suited for high performance systems. The reason for choosing C is its bare metal approach and integration into the Linux kernel with kernel modules.

3.2.2 Kernel Module Vs. Userspace Application

When sending data over the network, all calls to receive and send data have to go through the kernels internal network stack. These calls are usually in the form of system calls when running as a user-space application. Performing system calls (syscalls) can introduce a lot of performance overhead in the form of a context switch to kernel-space and copying data back and forth from user-space to kernel-space. A kernel module will be able to directly access the kernel functions, eliminating the overhead of system calls and the restricted access of userspace applications to the network stack.

Opting for a user-space application has the advantage of being cross-platform, meaning it can run on multiple operating systems without major modifications. In contrast, kernel modules are tightly bound to the Linux environment, which limits their usability to Linux hosts. This however is not a big problem as most servers are Linux hosts [1].

3.2.3 Connection splitting using Sockets

The PEP uses sockets to establish and split the connection between the client and the endpoint. Sockets provide a standardized way for programs to send and receive data over a network. They act as communication points, allowing for data exchange between them. For the PEPs purpose, sockets are ideal because they are configurable and come with dedicated buffers. This makes it easier to split and manage connections efficiently. Additionally, sockets are widely supported, ensuring compatibility and ease of integration [26].

A connection splitting proxy has the benefit of additionally splitting the connection into different domains. As discussed in Chapter 2, the internet consist of different domains with their own characteristics. Being able to split

the connection into different domains enables the PEP to adapt and select an appropriate congestion controllers based on the technology and topology of each domain.

Each socket connection can be configured to use a certain congestion controller, unrelated to the system wide default. This would be impossible with a single end to end connection. In addition to configuring the congestion controller, a socket can also add TCP options, such as TCP Fast Open, to a connection.

3.2.4 Bandwidth utilization

Another great benefit of connection splitting is potentially better bandwidth utilization. When using congestion control; the total bandwidth is usually determined by the lowest bandwidth on the path[17]. This is done to prevent overwhelming the link with the lowest bandwidth, as doing so would lead to congestion along the network path. Figure 3.3 shows a theoretical path between host A and host D; there are two links with high bandwidth, 60 Mbps and 75 Mbps. Because of the last link between C and D, the total bandwidth between A and D is 15 Mbps. This leads to a 25% bandwidth utilization of link AB and 20% utilization of link BC.

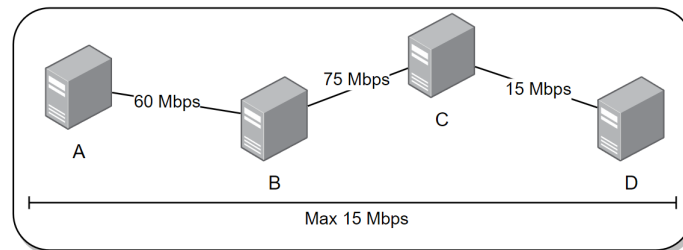


Figure 3.3: Example of poor bandwidth utilization.

Figure 3.4 shows the same path as Figure 3.3, however there is a connection splitting PEP at C. In that scenario we get two total bandwidths, one for each connection. The first connection between A and C has a total bandwidth of 60 Mbps, while the connection between C and D remains at 15 Mbps. Looking at the bandwidth utilization we now have 100% utilization on link AB, 80% utilization at BC. This is a major improvement from the original 25% and 20% utilization.

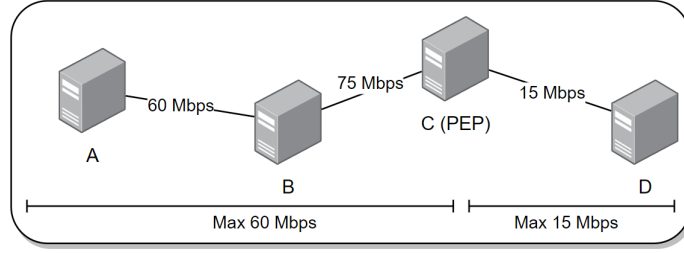


Figure 3.4: Example of good bandwidth utilization.

Link	Without PEP	With PEP
AB	25%	100%
BC	20%	80%
CD	100%	100%

Table 3.1: Table showing bandwidth utilization with and without a PEP.

Delay Table 3.1 gives an overview of the bandwidth improvements. However, even with the improvement of bandwidth utilization, we did not improve the end to end throughput from A to D as it is still limited by link CD. An important component is missing: Delay. Looking at the topology of Figure 3.4, we can assume an inherent lower delay between C and D, because of locality, than between A and D. If we now introduce a fluctuating bandwidth on the link CD the lower delay between C and D means the congestion controller can react faster.

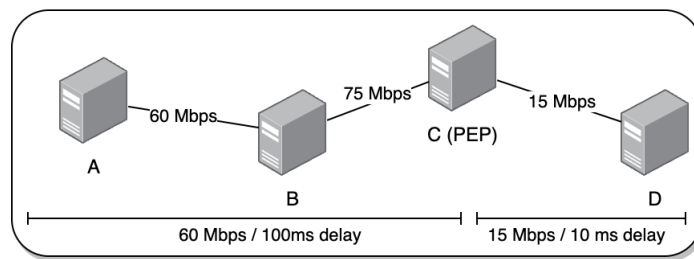


Figure 3.5: Split connection with bandwidth and delay.

Figure 3.5 introduces delay to the topology. Without the PEP we would have an end-to-end connection with 15 Mbps and 110 ms delay (220 ms RTT). Assuming the bandwidth on the link between C and D is highly fluctuating, the TCP congestion controller needs 220 ms (round trip) to react. However, if we split the connection at C, the time is reduced to 20 ms. This means the

congestion controller between C and D can react 1000% faster with a split connection, thereby better utilizing the fluctuating bandwidth.

3.2.5 PEP Selection

The connection process to the PEP, and later on the endpoint, is established by informing the PEP of the desired endpoint of the client. This can be achieved in a variety of ways where the goal is to attach additional information to the default connection process of TCP. Preferably, we do not want the overhead of needing an entire additional round trip just to pass this information. Figure 3.6 visualizes the RTT overhead of sending this information (in red) to the PEP before application data can be sent.

We do not want to change the normal socket based scheme of creating a connection either, as old applications would need to rewrite a lot of their code to adapt a new scheme. By providing a library for the client we make it can easy to integrate the PEP into existing environments. The library will provide a single function which mimics the default connection scheme. Implementation details are discussed in Chapter 4.

3.2.6 Connection Establishment

The PEP will attach data to the initial TCP handshake, this way we can inform the PEP of the endpoint without needing to inform it after an additional round trip time. Because of the ossified nature of TCP [12], changing the protocol itself is not an option. The realistic approach is reusing existing TCP functionality to append the desired data. Only a small amount of data, less than the usual Maximal Transmission Unit (MTU), is needed to inform the PEP of the endpoint.

TCP Options Since TCP Options can be attached to a TCP connection, a possibility would be to add a new TCP Option which would specify the endpoint. This TCP Option would need to be added by a kernel module, as it is not possible to to add custom TCP Options from user-space. This leads to another problem: how to specify from user-space that we wish to use the PEP.

A possible approach would be a socket option, using `setsockopt`. This however requires changes to the kernel, which raises the bar for adaptability. Another choice would be always attaching the TCP Options on connection

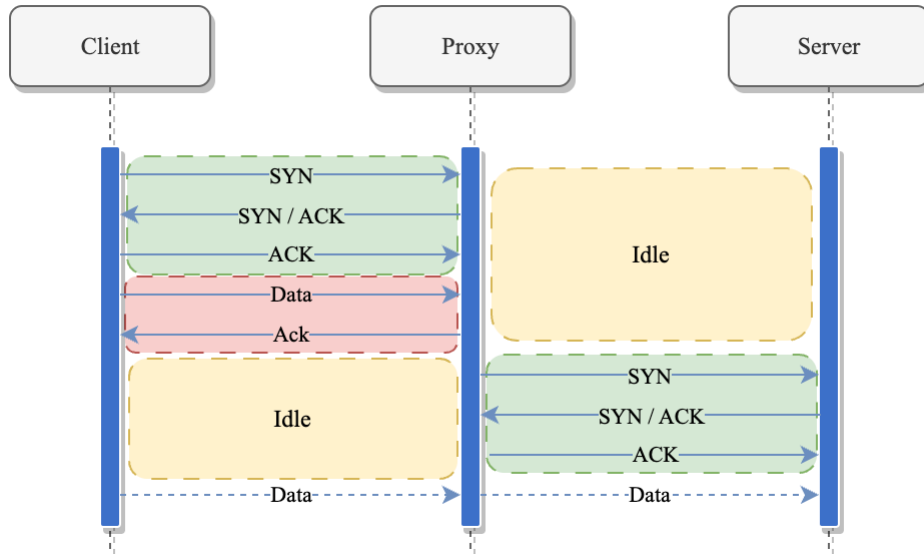


Figure 3.6: The TCP handshake procedure across PEP

addressed to a certain port such as 80/443. This however takes away the choice from the application, and makes it system wide instead.

Finally, another significant problem is that unknown TCP Options are often seen as a threat. Firewalls may drop the packets, or the options might be stripped by intermediate nodes [19]. This creates a challenge for the implementation and usability of the PEP. If the packets may be dropped because of our custom TCP options, then the PEP will only work in certain networks and scenarios. Although we only design a proof of concept, this is a trade off that is unlikely to pay off in the end [12].

TCP Fast Open Another possibility is using the existing TCP Fast Open option which can attach data to the initial TCP handshake. As discussed in the Chapter 2, using TCP Fast Open can reduce the amount of RTTs needed to establish a connection with both the PEP and endpoint. This requires the socket to be configured and enabled system-wide on the server machine. TCP Fast Open is also fault tolerant. Should either the PEP or the client not support TCP Fast Open, the meta-data will still be sent with the cost of an additional RTT.

Optimal Choice The most sustainable choice is TCP Fast Open, as adding new TCP Options is simply too unstable. Also, the goal of our PEP is not to

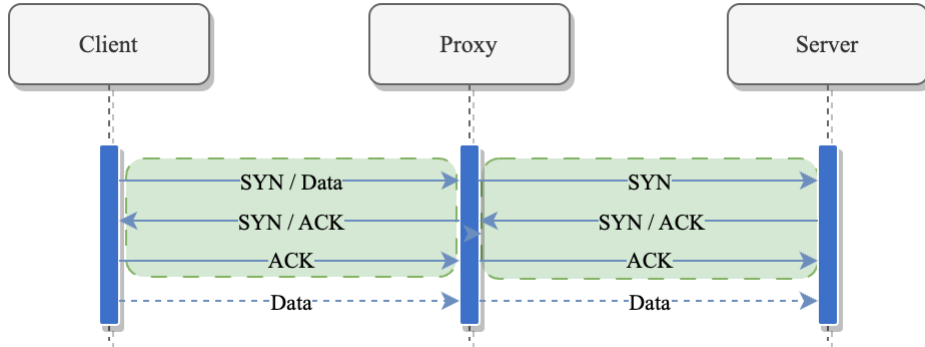


Figure 3.7: Optimal handshake across PEP (0 RTT)

change or extend TCP itself. Using TCP Fast Open also has the advantage of being able to add meta data in the size of an MTU, this enhances the future extensibility of the PEP.

Figure 3.7 shows the concept of the added data to the TCP handshake, reducing the idle time visualized in Figure 3.6.

3.3 Deployment (Transparency)

The deployment of the PEP is an important design aspect. In this context, deployment refers to the physical location of the PEP. The main factor which affects deployment is transparency; Chapter 2 briefly discussed the differences between the two approaches. This section will discuss the differences in the context of our PEP, especially in combination with wireless networks.

3.3.1 Transparent PEP

Transparent PEPs provide a deployment advantage since they can be integrated without altering the client or server. This means even older applications can utilize the PEP. Moreover, the PEP will inherently be on the same path as the original connection, which is a prerequisite for a PEP. But, if unknown middle boxes interfere with a connection, issues like lost TCP Options might occur. Additionally, if applications are not aware of the PEP, they can not adjust their behavior, restricting their future adaptability.

Most transparent connection splitting PEPs spoof the IP address of the original endpoint. From the perspective of the client, it is talking to the

actual endpoint. This may lead to issues, as several assumptions held by the client may no longer be valid. One assumption is that when a sender receives an ACK, the associated data segment was successfully received by the endpoint. But since the PEP is pretending to be the endpoint the ACKs only mean that the PEP received the data, not the endpoint. If the PEP would crash any data which the PEP has buffered would be lost. This would mean that the client assumes that the endpoint received the data, even though it did not.

3.3.2 Non Transparent PEP

In contrast, with a non-transparent PEP the deployment is more difficult. When a client needs to explicitly choose the PEP, it also needs to know that the PEP is on the path to the desired endpoint. Normally this would be extremely difficult to achieve. But, because the PEP is designed to be deployed at the base stations of wireless networks, such as 5G, we can assure that all traffic will pass through it. It is inherently on path, and eliminating the deployment problem of non-transparent PEPs.

A non-transparent PEP may still crash and buffered data may be lost. There is no assumptions by the client that the data was delivered to the endpoint. Clients are aware of the potential failures that may occur when using a PEP, similar to using a VPN. Non-transparent PEPs are more sustainable as they avoid the need to spoof IP addresses and include the clients which lets them adapt and avoid invalid assumptions.

Since applications are fully aware of the PEP when it is non-transparent, we can adjust the PEP's functionality based on the application using it. Applications can also adapt their behaviour, optimizing their operations based on the presence of the PEP, leading to additional performance improvements. Although side effects may still occur, applications are aware of them and can actively mitigate them.

This thesis will utilize a non-transparent PEP located at a base station, mostly for the benefits of future adaptability and the client being aware of the PEP. Being located at the base station assures that the PEP will be on path, which normally is a major challenge for non-transparent PEPs.

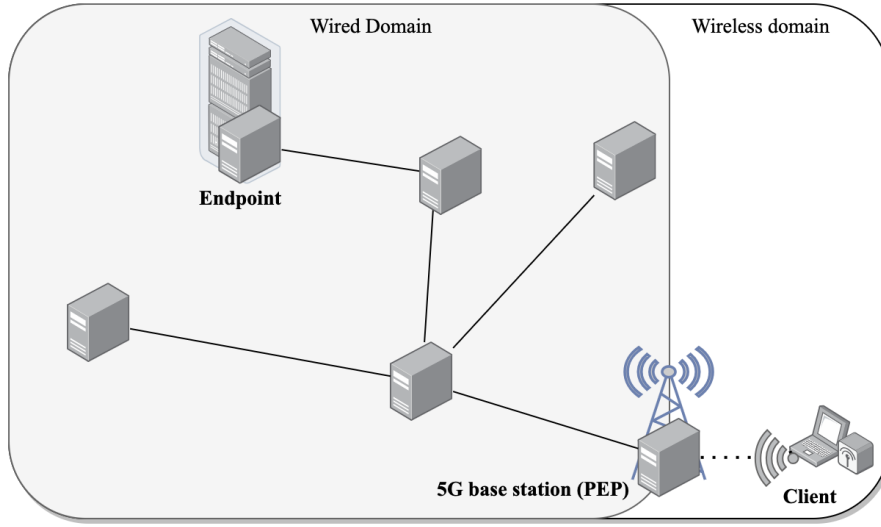


Figure 3.8: Deployment of a PEP at a base station

3.4 Buffering

Buffering is a core feature of the PEP. As mentioned before, the goal is to buffer non-interactive traffic packets at the PEP. Normally buffers in the network are NIC buffers or routing buffers. Those buffers affect all traffic passing through and will cause Bufferbloat if the buffers are too large. That's why, normally, the goal is to keep these buffers small to avoid disturbing interactive traffic. However, since the PEP is using sockets and splitting the connection, we can utilize a different buffer: the socket buffers.

3.4.1 Socket buffers

A socket buffer is a buffer specific for a particular socket. There are usually two buffers, one for read on the socket and one for write. This offers a great opportunity for buffering only a certain type of traffic, as a socket buffer does not affect traffic which flows through the host machine. A connection splitting proxy inherently has two sockets, one for the client and one for the endpoint. Which means we get a total of 4 buffers per 'flow', each individual read and write buffer can be configured for the best performance.

3.5 AQM & Scheduling

The ability to configure and utilize different scheduling algorithms is important for the success of the PEP. When buffering a lot of packets at the PEP we might clog up the sending queues of the host machine. This would result in high delays for the interactive traffic that has to share these queues with our buffered non-interactive data. To avoid this we utilize FQ-CoDel, as it assures a fair bandwidth for each flow.

FQ-CoDel usually preforms worse when looking at throughput compared to AQMs with bigger buffers such as BFIFO. This becomes especially clear with more fluctuating bandwidths, where buffering is important to avoid packet drops. Our PEP will utilize FQ-CoDel in combination with bigger buffers that will not affect other traffic, to both avoid unwanted delays and increase throughput.

This is another choice which will bind us to Linux, as achieving the same control and configuration on other operating system, such as Windows or MacOS, will be extremely difficult or even impossible. The impact of different scheduling algorithms will be shown and discussed in the Evaluation chapter.

3.6 Security

The PEP needs to assure that traffic flowing through the PEP does not get compromised. This, however, will not be a big concern as the PEP will only forward application data, meaning higher level security protocols such as TLS will still be end to end. This assures that the PEP can not read encrypted data or be used for malicious intents. The PEP can have the additional benefit of hiding the identity of the user from the endpoint, as only the PEPs connection will be seen by the endpoint.

3.7 Summary

In this chapter we have designed a high performance connection splitting PEP tailored for fluctuating bandwidth. We made the choice to keep it non-transparent to assure future adaptability and client configuration. Using sockets and their buffers, we avoid bufferbloat to not disrupt the interactive traffic. Finally, we use AQM and scheduling to avoid unnecessary delays on the outgoing links.

Implementation	0RTT	Connection Splitting	Special ACKs	Transparent	Large Buffers
MilliProxy		x	x	x	
PEP-DNA		x		x	
SnoopTCP			x	x	
Our PEP	x	x			x
TCP Convert	x	x			

Table 3.2: Table of design decisions based on different PEP implementations compared to ours.

Table 3.2 shows how our PEP compares with other PEPs based on features. We can see that only our PEP and 0-RTT TCP Convert support the 0-RTT TLV approach for metadata. Compared to most others our PEP is not transparent. This is an important distinction as non-transparent PEPs do not break the end to end semantics. The most important difference is the large buffers on our PEP, used only for non-interactive traffic. Although other PEPs also have larger than default buffers, our PEP has the largest configurable buffers, up to 2GB.

Chapter 4

Implementation

This chapter will explore the implementation of the TCP PEP, following up on the design choices made in the previous chapter. The development of the PEP will give a deeper understanding of the underlying mechanisms and how they aim to better utilize the 5G bandwidth. All aspects of kernel modules, PEP architectures and additional libraries will be covered.

4.1 Kernel Module

As mentioned in the design chapter, the PEP will be written as a kernel module instead of a user-space program. Running and creating a kernel module requires more initial preparation. Firstly, the biggest difference is that our PEP will run as a module inside the Linux kernel instead of as an application in its own virtual environment. Injecting a module into the Linux kernel is very different from simply running a binary.

A Linux kernel module is loaded and unloaded with the help of two functions that need to be defined:

Listing 4.1.1: The basic kernel module setup code.

```
1  /* Needed by all kernel modules */
2  #include <linux/module.h>
3  #include <linux/kernel.h>
4  #include <linux/init.h>
5
6  /* entry function */
7  static int __init onload(void) {
8      return 0;
9  }
10
11 /* exit function */
12 static void __exit onunload(void) {
13
14 }
15
16 module_init(onload);
17 module_exit(onunload);
```

Listing 4.1.1 shows a basic kernel module, defining and exporting the functions `onload()` and `onunload()`. The name of the functions bear no meaning, the important parts are the macros `__init`, `module_init` and `exit` macros[8]. When a kernel module is loaded, the function declared with `__init` is called.

Normally an application would terminate when it returns from its `main` function. Kernel modules however remain "loaded" when returning from the initialization function. This brings us to a new paradigm when programming: instead of having a running program, we install hooks and callbacks which change the default behavior of the kernel. A callback architecture can be less resource intensive as we do not need threads polling for data.

A kernel module is loaded by first compiling it into a `.ko` file and then loading with the `insmod` shell command:

```
$ insmod <module_name>.ko
$ rmmod <module_name>.ko
```

The `rmmod` command is used to unload a kernel module.

4.1.1 Kernel Hooks

The Linux kernel exposes many function tables and callbacks which designate what functions to call at certain events. Most drivers and congestion

controllers are implemented in this manner. A predefined struct is allocated and populated with custom functions, and installed with an existing kernel function.

Listing 4.1.2: Example of a TCP congestion controller module

```
1
2 static void my_init(struct sock* sk);
3 static u32 my_ssthresh(struct sock* sk);
4 ...
5
6 static struct tcp_congestion_ops mycc __read_mostly = {
7     .init          = my_init,
8     .ssthresh       = my_ssthresh,
9     .cong_avoid     = ...,
10    .set_state      = ...,
11    .undo_cwnd      = ...,
12    .pkts_acked     = ...,
13    .owner          = THIS_MODULE,
14    .name           = "tuner",
15 };
16
17 /* entry function */
18 static int __init onload(void) {
19     return tcp_register_congestion_control(&mycc);
20 }
21
22 /* exit function */
23 static void __exit onunload(void) {
24     tcp_unregister_congestion_control(&mycc);
25 }
26
27 module_init(onload);
28 module_exit(onunload);
```

In Listing 4.1.2 we demonstrate how a TCP congestion controller is implemented, in the context of a kernel module. This allows run-time modification of kernel behaviour, and is a programming paradigm that will be useful when implementing the PEP.

4.1.2 Linux Version and Distribution

The kernel module was developed for Linux Kernel > 5.10, more specifically Debian 11 (Bullseye). Older versions of Linux might encounter problems as some of the kernel functions change between kernel versions. The choice of

kernel version dictates what functions are available to kernel modules. Usually helper functions are the most volatile, so avoiding those is an important goal for compatibility.

4.2 TLV Library

Regarding the endpoint addressing and selection by the client, a custom shared library has been implemented. The goal is to keep the client code as simple and close to its original form as possible, but still be able to communicate the desired endpoint, options and meta-data to the PEP. Additionally, we want to send this information by using TCP Fast Open which has a similar connection routine as default sockets do.

Normally a socket would first `create`, `connect` and then `send` data using the `send` system call. The creation of a socket is the same for both with and without TCP Fast Open. This leaves us with connection as the main problem. The way we connect to a host using TCP Fast Open is by skipping `connect` and instantly jump to `sendto`. As we can see from the socket system calls, both `connect` and `sendto` take a `struct sockaddr` as a parameter.

A `struct sockaddr` defines the endpoint to which one wants to connect. `sendto` has a flag option which allows to configure how messages are sent. If we supply the `MSG_FASTOPEN` flag, `sendto` will automatically connect and deliver the given message within the handshake. Subsequent uses of `send` will function as if the `connect` function was used.

Important socket system calls

```
int connect(
    int sockfd,
    const struct sockaddr *addr,
    socklen_t addrlen
);

ssize_t send(
    int sockfd,
    const void *buf,
    size_t len,
    int flags
);

ssize_t sendto(
    int sockfd,
    const void *buf,
    size_t len,
    int flags,
    const struct sockaddr *dest_addr,
    socklen_t addrlen
);
```

4.2.1 Custom connect function

The library will replace the original `connect` with a custom implementation `pep_connect`. The original `connect` and `sendto` have a lot of parameters in common, specifically the `const struct sockaddr` which is used to identify an endpoint. In the context of the PEP, this would identify the final endpoint to which the client wants to connect. The goal of the PEP connect function is to replace the `sockaddr` given by the client with one that identifies the PEP instead, but still forwards the original `sockaddr` to the PEP to establish the proxy connection.

The signature of our custom implementation mimics the original `connect`. The main difference is the addition of a `flags` parameter for easier customization. Inside our custom function we allocate space for a new `struct sockaddr` which we will fill with the IP address and port of the PEP, while at the same time creating a message with original IP address and port of the endpoint. Finally we call `sendto` with our message and the new `struct sockaddr`, using `MSG_FASTOPEN` to both connect and deliver the message.

The custom connect function signature

```
int pep_connect(  
    int sockfd,  
    const struct sockaddr* addr,  
    socklen_t len,  
    int flags  
);
```

4.2.2 TLV

A good choice for sending options and meta-data is in the form of TLVs, formally known as Type–Length–Value options. The idea is that all options can be defined by a Type, Length and Value. The **type** defines the type of an option, which types exist and what they mean are up to the users of the library to decide. Common types are "Version", "Error", etc. Adding new types is very easy and requires little modification. The TLV is practically implemented as its own message, but can also be appended at the start of a transmission like a header.

TLV structures

```
struct __tlv_header {  
    unsigned char version;  
    unsigned char len;  
    unsigned short magic;  
};  
  
struct tlv {  
    unsigned char type;  
    unsigned char length;  
    unsigned short value;  
    unsigned int optional;  
};
```

TLV Implementation In our design, TLVs are structured as a continuous buffer, consisting of a TLV header followed by subsequent options. This header provides details on the version, the number of options, and a unique magic number for validation. The **type** spans one byte, which means we limit ourselves to 255 possible types. The current implementation only uses six, which means we have enough space for future extensions. The **length** variable is mainly used to indicate an optional data segment called **optional**. The size of the default value is two bytes with an additional option of size

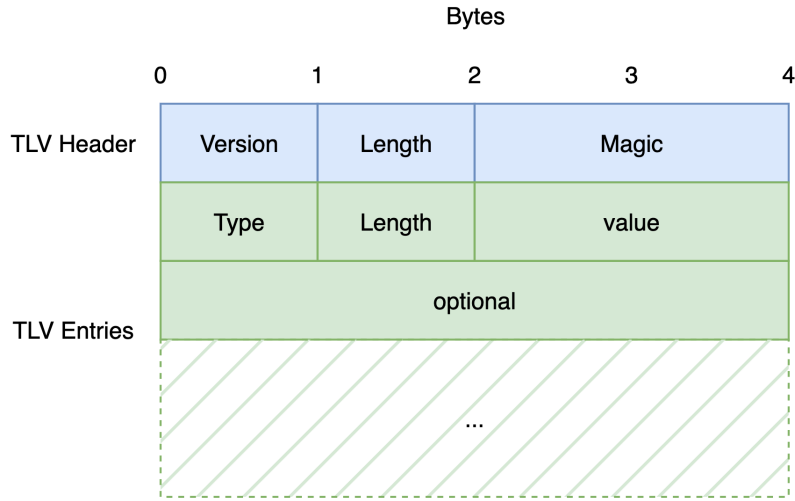


Figure 4.1: Example of TLV header and entries

four bytes. All together the struct uses eight bytes or 64 bits, which mostly comes from the fact that we need at least six bytes for the IP address and port alone.

Problems using TFO When using TCP Fast Open, the `sendto` function instantly returns when the message has been sent with no confirmation of a successful connection. This is in contrast to `connect` which only returns when a connection was successfully established and an error has occurred. In practice its not a big problem as any calls to `send` will either wait for the connection to be established or return an error if the connection was unsuccessful. Only applications which assume a connection is successful by the `connect` function alone and use no subsequent `send` or `recv` might encounter problems. Those applications might need further adjustment to be able to utilize the PEP.

4.2.3 TLV Options

The options of the TLVs define the functionalities a PEP can provide, for our PEP we only need six options. The basic `info` and `error` options are included alongside some information about the TCP connection such as extended headers. The most important type is `connect`, it specifies a port and IP address and is used to communicate the endpoint to the PEP. A TLV message can include as many options as a client desires.

```
enum __tlv_types {
```

```

    TLV_INFO = 0x1,          // Info TLV
    TLV_CONNECT = 0xA,       // Connect TLV
    TLV_EXT_TCP = 0x14,      // Extended TCP header
    TLV_SUPP_EXT = 0x15,     // Supported TCP extension
    TLV_COOKIE = 0x16,       // Cookie TLV
    TLV_ERROR = 0x1E         // Error TLV
};

```

4.2.4 Shared Library

The PEP TLVs are implemented as a shared library which is both used by the PEP itself and applications. The applications will use the library to create TLVs for connecting to the PEP, while the PEP uses the library to validate and read the TLV options. A shared library can be created by passing the `-shared` flag to the linker.

```
$ gcc file.c <path>/<lib>.so -o file.o
```

4.3 PEP - Internals

The internals of the PEP will consist of many important components. Importantly, we have the main server socket, socket pairs (tunnels) and deferred works. The PEP itself will need to keep track of the PEP state and the state of all the tunnels, additionally we need to keep track of all running tasks. Since our code will be a kernel module and run 'inside' the kernel, we will have access to a lot of existing infrastructure which normally is only accessible by the kernel.

4.3.1 Architecture

The PEP architecture is mainly based on a server socket and multiple socket tunnels. The server socket acts as the entry point of the PEP, clients will connect to the server socket which will in turn create a new socket and connect to the endpoint. The server consists of a `pep_state` which holds the server socket, work queues and a list of tunnels. The server socket accepts new clients and reads their TLV header, if no header is present the connection will be closed. After successfully reading the TLV, a new socket is created and a connection is attempted to the endpoint specified by the TLVs.

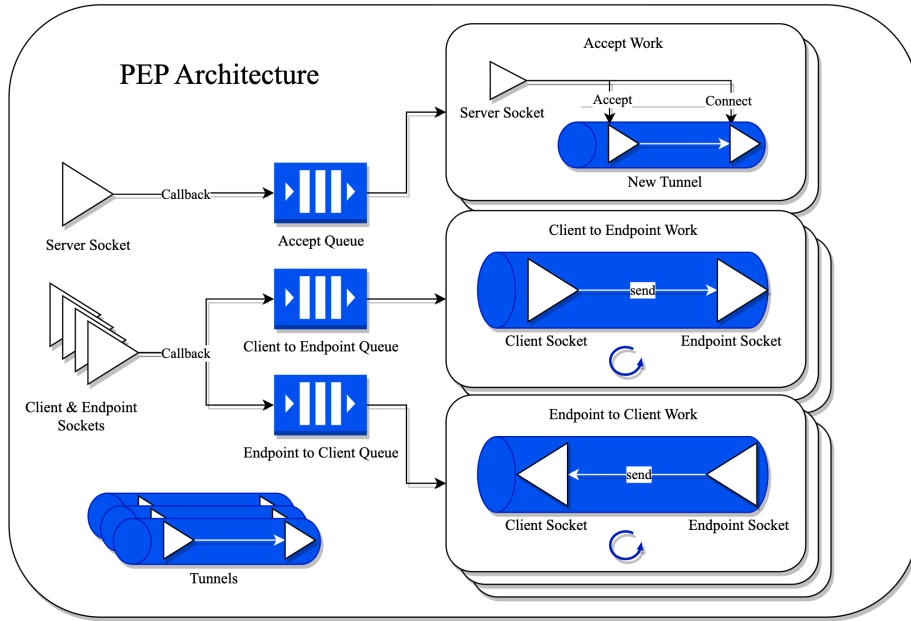


Figure 4.2: The architecture of the PEP

```

struct pep_state {
    atomic_t state;

    struct pep_state_ops* ops;
    struct pep_state_work_ops* work_ops;

    struct socket* server_socket;
    struct workqueue_struct* accept_wq;
    struct workqueue_struct* forward_c2e_wq;
    struct workqueue_struct* forward_e2c_wq;
    struct work_struct accept_work;
    struct list_head tunnels;
    unsigned int total_tunnels;
};

```

Figure 4.3: PEP State Structure Code

Tunnels After the endpoint connection is established, the PEP will create a **tunnel**. A **tunnel** consists of two sockets, where data is transferred between them. Additionally to the sockets, a **tunnel** also holds the information about the sockets and the proxy connection, this includes total amount of

data transferred.

PEP Tunnel Structure

```
struct pep_tunnel {
    unsigned int id;
    struct pep_connection client;
    struct pep_connection endpoint;
    int total_client;
    int total_endpoint;
    struct work_struct c2e;
    struct work_struct e2c;
    struct list_head list;
    struct pep_state* server;
    int state;
    int recv_callbacks;
    int packets_fowarded;
};
```

4.3.2 Kernel Sockets

Normally an application would interact with socket descriptors when sending and receiving data. However, inside the kernel we have access to the actual sock and socket structures. This gives us full access to all functionality and meta-data a socket can offer, such as `sk_buff` queues, callbacks and statistics like `sk_drops`. An `sk_buff`, which stands for socket buffer, is the kernel struct that holds the information about a packet. It includes both the headers and data and is created when the kernel receives a packet. After that it is routed to the corresponding socket.

```

struct sock {
    atomic_t sk_drops;
    void* sk_user_data;
    ...
    struct sk_buff_head sk_error_queue;
    struct sk_buff_head sk_receive_queue;
    ...
    struct socket *sk_socket;
    ...
    void (*sk_state_change)(struct sock *sk);
    void (*sk_data_ready)(struct sock *sk);
    void (*sk_write_space)(struct sock *sk);
    void (*sk_error_report)(struct sock *sk);
    ...
};

```

Figure 4.4: Kernel Socket Structure

Socket Operations The kernel socket struct includes a `struct proto_ops` member which holds usual socket operations such as `accept`, `connect` and `bind`. Figure 4.5 shows an overview over the most interesting socket operations. Most notably, the `setsockopt` operation is important as it allows the configuration of socket options such as `TCP_FASTOPEN`. Additionally, it can configure the socket buffer sizes using `SO_SNDBUFSIZE` and `SO_RCVBUFSIZE` which is an important aspect of the PEP.

Socket Callbacks Especially of interest is the callback function `sk_data_ready`. It is called whenever an `sk_buff` is received by a socket. That is the case for any kind of packet, not only user data but also includes the TCP handshake packets. This callback can be overwritten by a custom implementation, which is useful to detect any new data on the socket. For example one could inspect or manipulate headers before calling the original `sk_data_ready` function. All callbacks take the socket as a parameter, thereby giving access to the socket state.

Socket User Data Kernel sockets allow modules to add extra information to a specific socket. The pointer `sk_user_data` can be used to point to any data defined by a module. This in combination with the callbacks is a powerful tool for customizing socket behavior. This allows us to add additional

state information to the socket. In our case we will use `sk_user_data` to store the server stat.

```
struct proto_ops {
    int family;
    struct module *owner;
    int (*release)(struct socket *sock);
    int (*bind)(struct socket *sock,
                struct sockaddr *myaddr,
                int sockaddr_len);
    int (*connect) (struct socket *sock,
                   struct sockaddr *vaddr,
                   int sockaddr_len, int flags);
    int (*accept) (struct socket *sock,
                  struct socket *newsock, int flags, bool kern);
    ...
    int (*listen) (struct socket *sock, int len);
    int (*shutdown) (struct socket *sock, int flags);
    int (*setsockopt)(struct socket *sock, int level,
                     int optname, sockptr_t optval,
                     unsigned int optlen);
    int (*getsockopt)(struct socket *sock, int level,
                     int optname, char __user *optval, int __user *optlen);
    int (*sendmsg)(struct socket *sock, struct msghdr *m,
                  size_t total_len);
    int (*recvmsg) (struct socket *sock, struct msghdr *m,
                    ...
};
```

Figure 4.5: Socket Protocol Operations from Linux - net.h [18]

4.3.3 Work Queues

To handle multiple concurrent events, the PEP uses **work queues**. A work in the Linux kernel is a way of handling kernel threads. They allow for a more reactive approach to threads, as a work can be queued on demand and execute simple tasks. A work is usually queued into a **work queue**. Each **work queue** represents a task and works which are queued will wait in the queue till they can be run. A task is defined as a function, to which context can be added through a **work** parameter.

```

struct work_struct {
    atomic_long_t data;
    struct list_head entry;
    work_func_t func;
#ifdef CONFIG_LOCKDEP
    struct lockdep_map lockdep_map;
#endif
};

```

Figure 4.6: Work struct from Linux - workqueue.h

4.3.4 Works

A work is defined by `work_struct` (see Figure 4.6) which holds information associated with the work, most importantly the `work_func_t func`. `func` must be a pointer to a function which takes a `struct work_struct` as parameter. This is the function which will be called when a work is scheduled. The work structure should be created by the user and not be allocated on the stack as external services need access to it. Normally the work structure will be part of another structure as it is in Figure 4.3.

Listing 4.3.1: Work initialization example

```

1 void my_work_handler(struct work_struct *work);
2
3 struct work_struct my_work;
4 struct workqueue_struct * my_workqueue;
5
6 my_workqueue = create_singlethread_workqueue("workqueue")
7 ;
8 INIT_WORK(&my_work, my_work_handler);
9 queue_work(my_workqueue, &my_work);

```

Work state problem Initiating a Work poses a challenge: how to effectively track the state. In this context, 'state' refers to the status of a PEP tunnel, which includes the sockets and a reference to the server. A simple approach is to store the PEP server in a global variable. However, not only is this poor coding practice, but it also doesn't solve how to get the sockets. A function that transfers between two sockets must be aware of the specific

sockets to use. The work alone doesn't provide this context.

To solve this issue we can make use of a macro which Linux provides. `containerof` is a macro that retrieves a reference to the parent structure of any given struct. This, in combination with the fact that we get the original work struct as a parameter, allows us to access the tunnel state if we attach the work to the tunnel struct.

Listing 4.3.2: Work using `containerof` example

```
1 struct parent {
2     struct work_struct my_work;
3     // ...
4 };
5
6 void my_work_handler(struct work_struct *work)
7 {
8     struct parent * my_parent;
9
10    my_parent = container_of(work, struct parent, my_work);
11    // ...
12 }
```

As described in the PEP Tunnel Structure (see above struct 4.3), the PEP has 3 main work queues:

```
struct workqueue_struct* accept_wq;
struct workqueue_struct* forward_c2e_wq;
struct workqueue_struct* forward_e2c_wq;
```

Accept Work Queue The main work queue for the server is the accept work queue. When a client attempts to establish a connection to the PEP, we want to receive that notification and queue an accept work. This is an alternative to a thread blocking on accept. The notification can be achieved by replacing the `sk_data_ready` with a custom function which checks the TCP state and queues an accept work.

Listing 4.3.3: Accept callback function

```
1 void pep_listen_data_ready(struct sock* sk)
2 {
3     struct pep_state* server;
4
5     read_lock_bh(&sk->sk_callback_lock);
6     server = sk->sk_user_data;
7
8     /* Queue accept work */
9     if(sk->sk_state == TCP_LISTEN){
10         queue_work(server->accept_wq, &server->accept_work);
11     }
12     read_unlock_bh(&sk->sk_callback_lock);
13
14     default_data_ready(sk);
15 }
```

The function `pep_listen_data_ready` (see Listing 4.3.3) outlines the process of queuing an accept work. We retrieve the PEP server state from the socket `sk_user_data` variable, afterwards we check the socket state for `TCP_LISTEN` which indicates that the socket is ready to accept a connection. If the socket has the correct state we queue the `accept_work` on the `accept_wq` work queue, which is part of the PEP server.

Packet Forwarding Queues The second two work queues handle forwarding of packets, one for each direction. The reason we use two separate works and work queues is because the function has to identify which socket it should read from and which one it should send to within the tunnel socket pair.

Listing 4.3.4: Forwarding callback function

```
1 void pep_client_data_ready(struct sock* sk)
2 {
3     struct pep_tunnel* tunnel = sk->sk_user_data;
4     tunnel->recv_callbacks++;
5
6     queue_work(tunnel->server->forward_c2e_wq, &tunnel->c2e
7 );
8     default_data_ready(sk);
9 }
```

4.3.5 Kernel TCP receive and send

In the kernel, the message functions `recvmsg` and `sendmsg` are used for reading from and sending data to sockets. However, these functions bring some overhead that can lead to code clutter. To reduce this, two helper functions will be utilized instead, `pep_tcp_receive` and `pep_tcp_send`. These functions mimic the usage of the `send` and `recv` system calls, while abstracting away the complexity of the message functions.

```
int pep_tcp_receive(struct socket *sock, u8* buffer, u32 size);
int pep_tcp_send(struct socket *sock, u8* buffer, u32 size);
```

The custom functions only take in the socket, buffer and size of buffer as parameters. These functions also abstract away certain error handling which are common in the kernel space such as `EAGAIN` and `ERESTARTSYS`, both of which indicate to retry a function.

4.4 PEP - Server

4.4.1 Creation

The creation of a PEP server is achieved by the `pep_new_server` function. The function is responsible to allocate memory for the server and attach the default operations, work functions and callbacks. These default values can be modified and overwritten for custom functionality, how to extended and modify the PEPs behavior will be explored later.

Listing 4.4.1: PEP server creation

```
1 struct pep_state* pep_new_server(void)
2 {
3     struct pep_state* server = kzalloc(sizeof(struct
4     pep_state), GFP_KERNEL);
5     if(!server){
6         return NULL;
7     }
8     server->ops = &default_server_ops;
9     server->work_ops = &default_server_work_ops;
10    server->callbacks = &default_server_callbacks;
11
12    return server;
13 }
```

4.4.2 Server initialization

Server state initialization consists of creating and configuring the main server socket and work queues. The server is responsible for accepting clients and creating pep tunnels. Additionally, it holds the work queues for all the PEPs functionality. Socket configuration consists of replacing the `sk_data_ready` data callback, setting `sk_user_data` to the server itself and using `setsockopt` to set both `TCP_FASTOPEN` and `TCP_NODELAY`. `TCP_NODELAY` is set to avoid socket latency by 'waiting' for larger frames.

Listing 4.4.2: PEP server initialization (Simplified)

```

1  int pep_server_init(struct pep_state* server, u16 port)
2  {
3      ...
4
5      /* socket creation */
6      struct sock* sk = NULL;
7      struct sockaddr_in saddr;
8      ret = sock_create_kern(&init_net, ..., &sock);
9      if(ret){
10         printk(KERN_INFO "[PEP] init_core: Error creating
            socket\n");
11         return -EPEP_GENERIC;
12     }
13
14     ...
15     server->state = ((atomic_t){(PEP_SERVER_RUNNING)});
16
17     /* use our own data ready function */
18     write_lock_bh(&sk->sk_callback_lock);
19     sk->sk_user_data = server;
20     sk->sk_data_ready = server->callbacks->
        server_data_ready;
21     write_unlock_bh(&sk->sk_callback_lock);
22
23     /* pep server connection info */
24     ...
25
26     pep_setsockopt(sock, TCP_FASTOPEN, 5);
27     pep_setsockopt(sock, TCP_NODELAY, 1);
28
29     ... bind and listen ...
30
31     server->accept_wq = alloc_workqueue("accept_wq",
        WQ_HIGHPRI|WQ_UNBOUND, 0);
32     server->forward_c2e_wq = alloc_workqueue("c2e_wq",
        WQ_HIGHPRI|WQ_UNBOUND, 0);
33     server->forward_e2c_wq = alloc_workqueue("e2c_wq",
        WQ_HIGHPRI|WQ_UNBOUND, 0);
34
35     ...
36
37     return 0;
38 }

```

Line 20 in Listing 4.4.2 configures the `sk_data_ready` callback to the

aforementioned `pep_listen_data_ready`. When overwriting the `sk_data_ready` we need to make sure we hold the socket `sk_callback_lock` to avoid any race conditions, this is done in line 17 and 21. The accept and forward work queues are allocated and created with `WQ_HIGHPRI` and `WQ_UNBOUND`.

`WQ_HIGHPRI`

Work items of a highpri wq are queued
to the highpri worker-pool of the target cpu.

`WQ_UNBOUND`

Work items queued to an unbound wq are served
by the special worker-pools which host workers
which are not bound to any specific CPU.

Figure 4.7: <https://docs.kernel.org/core-api/workqueue.html>

The reason for both `WQ_HIGHPRI` and `WQ_UNBOUND` is to avoid any added latency by work queuing, especially if there are a lot of other works being queued. Work queues are used by the kernel for any deferred work, which means that there might be competition for both CPU and scheduling. The before mentioned flags assure that the PEP work's are prioritized.

Left out from Listing 4.4.2 is the creating of the accept work itself. It uses the `INIT_WORK` macro (shown in 4.3.1) with the `pep_listen_data_ready` function. After that the server is configured and ready for accept callbacks.

4.4.3 Accept and Endpoint connection

The accept work will call the `pep_server_accept_work` which is responsible for creating a new tunnel and connection to the desired endpoint. Firstly, the server state is fetched by using `container_of`, after which we assert that the server is in a operational state. Next, the kernel will accept a new connection in a non-blocking fashion as we know there is an incoming connection request.

Listing 4.4.3: PEP server accept function (Simplified)

```
1 int pep_server_accept_work(struct work_struct *work)
2 {
3     struct pep_state* server = container_of(work, struct
4     pep_state, accept_work);
5
6     rc = kernel_accept(server->server_socket, ...);
7
8     ... read data from socket ...
9
10    /* Validate tlv header. */
11    if(!tlv_validate(buffer)){
12        return;
13    }
14
15    /* Get connect tlv options from tlv buffer */
16    tlv = tlv_get_option(TLV_CONNECT, buffer);
17    if(tlv == NULL || tlv->length != 6){
18        sock_release(client);
19        return;
20    }
21
22    endpoint = pep_endpoint_connect(tlv->optional, tlv->
23    value);
24    if(NULL == endpoint){
25        sock_release(client);
26        return;
27    }
28
29    ... configure sockets and tunnel ...
30
31    return 0;
32 }
```

After successfully accepting the client we immediately allocate a buffer and read from the client. We expect it to send a TLV with TCP Fast Open, so we use the TLV library to validate and read the TLV options. Note that even if TCP Fast Open should fail, the connection can still be established at the cost of an additional round trip time. Specifically, we look for the `TLV_CONNECT` option, which will be used to connect to the endpoint.

4.4.4 Multiple Servers

Because of the callback nature of the PEP we can create multiple servers on the same host machine. By not having a 'global' server in the kernel module, we are able to potentially create as many servers as are desirable. The user only has to keep track of the server pointers, while the implementation of the server keeps track of the states and correct callback handling through the works. This means that each server may use the same callback function, but the state will vary.

4.5 PEP - Clients

4.5.1 Client Sockets - Endpoint Sockets

As discussed in Section 4.3.4, there are two functions responsible for forwarding packets. After queuing a forwarding work from a callback, the corresponding function is executed.

Listing 4.5.1: Client Forwarding Function (Simplified)

```
1 void pep_client_receive_work(struct work_struct *work)
2 {
3     int ret = 1;
4     int ret_forward;
5     struct pep_tunnel* tun = container_of(work, struct
        pep_tunnel, c2e);
6
7     unsigned char *buffer = kzalloc(...);
8     if (!buffer) {
9         return;
10    }
11
12    while(ret > 0){
13        ret = pep_tcp_receive(tun->client.sock, ...);
14        if(ret > 0){
15            ret_forward = pep_tcp_send(tun->endpoint.sock, ...)
16            ;
17            tun->total_client += ret_forward;
18            tun->packets_fowarded++;
19        } else {
20            if(pep_tunnel_is_disconnected(tun)){
21                pep_tunnel_close(tun);
22                return;
23            }
24        }
25        kfree(buffer);
26    }
```

Listing 4.5.1 shows the function that forwards data from the client to an endpoint. Firstly, we retrieve the tunnel state by using `container_of`, this gives us the socket which triggered the original callback. After that, a buffer is allocated and data is read from the client and forwarded to the endpoint. This function will also run while there is data to send, the reason is that it is more effective to read all the data that is available than to wait for a callback and work queue to trigger.

If a socket returns 0 or less we check if the connection is closed, that is because the closing of a connection will trigger the same `sk_data_ready` callback. However, when a socket is closed it will return 0 or an appropriate error code.

Socket Shutdown It is important to properly handle the shutdown procedures of sockets because the PEP will manage two sockets per tunnel. An incorrect shutdown of a socket will result in loss of data or even worse a kernel panic. When the PEP detects that either the client or endpoint has closed their socket, we will track that event. But we cannot fully close and shutdown the sockets before we have read all the data from it. When all data has been transferred, and the sockets are in the `TCP_CLOSE` or `TCP_CLOSE_WAIT` state, we will shutdown the sockets using `kernel_sock_shutdown` with `SHUT_RDWR` to shutdown both read and write.

4.5.2 PEP Connections

By design the PEP is able to handle multiple connections at once. Each socket pair has its own work structure for client - endpoint and endpoint - client communication. This means each work structure can run in parallel, in both directions. Each tunnel (socket pair) is added to a linked list in the server state. This allows the PEP to have access to them in case we need to prematurely terminate the connections. This will avoid any memory leaks since we manage the memory for the tunnels.

4.5.3 Module Customization

The PEP will follow a similar approach as congestion control modules when it comes to how the PEP is configured. The basic accept and forward functions will be defined by a table, which can be created by any future module. Each server has a pointer to this table, which it uses when creating works. This mimics the way we configure socket callbacks and makes each PEP server more customizable. Each individual PEP server can have different forward functions, or keep the original ones.

```
struct pep_state_work_ops {
    void (*accept)(struct work_struct *work);
    void (*forward_c2e)(struct work_struct *work);
    void (*forward_e2c)(struct work_struct *work);
};
```

Figure 4.8: Work operation table

The work function table in combination with the fact that the PEP supports multiple servers, means that each individual server can be configured differently without needing to change any of the original code. Simply creating a new `struct pep_state_work_ops` and supplying new work functions is enough.

```
struct pep_socket_callbacks {
    void (*server_data_ready)(struct sock* sk);
    void (*client_data_ready)(struct sock* sk);
    void (*endpoint_data_ready)(struct sock* sk);
};
```

Figure 4.9: Callback function table

4.5.4 System Configurations

The PEP will require some configuration outside of the kernel module itself. Linux uses `sysctl` for system configuration. Most importantly we want to enable TCP Fast Open and IP forwarding; IP forwarding will allow the Linux machine to act as a router and forward packets. Both these options are under `net.ipv4`:

```
$ sysctl -w net.ipv4.tcp_fastopen=3
$ sysctl -w net.ipv4.ip_forward=1
```

Buffer sizes The PEP works by buffering as much data as possible on the sockets themselves. This way we avoid buffering interactive traffic that simply passes by. The size of a socket's buffer can be configured with `setsockopt`, however this is not reliable and has to be done for each socket. Instead we can configure the socket buffer sizes system-wide. Under `net.core` there exist configurations for overall receive buffer sizes: `rmem_max` and `wmem_max`. There also exists the same for the default variables `rmem_default` and `wmem_default`.

```
$ sudo sysctl -w net.core.rmem_max=<size>;
$ sudo sysctl -w net.core.wmem_max=<size>;
$ sudo sysctl -w net.core.rmem_default=<size>;
$ sudo sysctl -w net.core.wmem_default=<size>;
```

Additionally, under `net.ipv4` there are options to configure the amount of memory in bytes a TCP socket can buffer for both total, read and write. Each contains three numbers: the minimum, default, and maximum values.

```
$ sudo sysctl -w net.ipv4.tcp_rmem='<min size> <size> < max size>';  
$ sudo sysctl -w net.ipv4.tcp_wmem='<min size> <size> < max size>';  
$ sudo sysctl -w net.ipv4.tcp_mem='<min size> <size> < max size>';
```

4.5.5 Userspace

The PEP is designed as a kernel module, however the functionality is not dependent on any kernel functionality. Userspace alternatives to work queues and callbacks exist, such as threads. The PEP can be rewritten with these alternatives. Theoretically the PEP can run as a userspace program for higher compatibility at the cost of performance loss.

4.6 Memory

Because we chose C as our programming language of choice, we are responsible for any memory allocated on the heap. The main places where dynamic memory is allocated is the server itself, tunnels and buffers for data forwarding. Both the tunnels and buffers are managed by the PEP server, so the memory will be freed when the server is terminated. This leaves the server to be freed, which is left up to the user. The easiest solution is to use the `pep_server_clean` function in the `__exit` module function.

Beside direct memory allocations, the PEP also needs to close and free sockets, works and work queues. These are all handled by the `pep_server_clean` function. Memory management is especially important as the PEP is a kernel module, which means any memory that is not freed will affect the entire system. Additionally, any improper handling of the memory, such as double free's or using freed memory will result in a kernel panic.

Chapter 5

Evaluation

Our evaluation of the PEP centers on measuring the completion times for non-interactive traffic. We aim to demonstrate how the PEP can speed up this traffic compared to standard end-to-end methods. It's important that this enhancement does not negatively affect interactive traffic. To illustrate this, we'll use graphs to show and compare the PEPs impact on interactive traffic.

5.1 Traffic Control Options

Linux has support for network interface configurations using the TC (traffic control) command. TC allows the configuration of packet scheduler, bandwidth, delay and jitter. These options combined with the fact that each network interface can have its own configuration, allows for very precise testing environments.

5.2 Ocarina Testbed

The Testbed consists of four hosts, the first is the controller from which we orchestrate the experiment. The last three represent the sender, router and receiver. The controller is connected to the other hosts through a management network, while each of the other hosts are directly connected to each other using 10G Ethernet controllers. All tests are run using physical hosts to get the most realistic results and avoid unexpected side effects of virtual machines.

Figure 5.1 shows a technical overview of the testbed configuration.

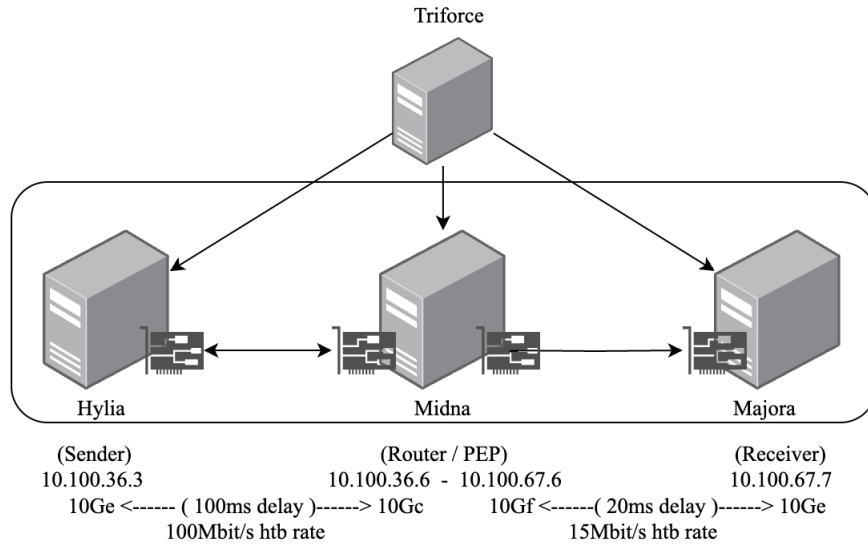


Figure 5.1: Overview of the Ocarina testbed

5.2.1 Hosts and Hardware

All three hosts used in the testbed run Debian 11 (bullseye) with an Intel Xeon E5-2620 v4 processor, which features 32 cores at 3.0 GHz. Additionally the hosts have 64 GB of RAM each. The choice of hardware should eliminate any potential host congestion [2] and assure that the tests are not affected by old or dated hardware.

_,met\$\$\$\$\$gg.	ocarina@hylia
,g\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$P.	OS: Debian 11 bullseye
,g\$\$\$P"" ""Y\$\$.	Kernel: x86_64 Linux 5.10.0-26-amd64
,\$\$\$' '\$\$.	Uptime: 76d 12h 21m
',\$\$\$,ggs. '\$\$b:	Packages: 1586
'd\$\$' ,P"' . \$\$\$	Shell: bash 5.1.4
\$\$\$ d\$' , \$\$\$	Disk: 78G / 496G (17%)
\$\$: \$\$ - ,d\$\$'	CPU: Intel Xeon E5-2620 v4 @ 32x 3GHz
\$\$\; Y\$b._ _ ,dP'	GPU: Matrox Electronics Systems Ltd.
Y\$\$. '."Y\$\$\$\$P"	RAM: 1735MiB / 64303MiB
'\$\$b "-._	
'Y\$\$	
'Y\$\$.	
'\$\$b.	
'Y\$\$b.	
'"Y\$b._	
'""	

Figure 5.2: Screenfetch results

It is important to note that the PEP will run on a normal Linux host which acts as a router, and not on an actual router. Using a normal Linux host may introduce small overheads as the system will support user interactions and a lot more sub systems which usually are not present on a pure router.

5.3 Initial test

To evaluate the initial performance of the PEP we use a simple topology consisting of a sender, router and receiver. The router will act as our hypothetical base station at which the PEP resides. And the connection between the router and the receiver will act as the unstable fluctuating bandwidth domain.

Sender — Router — Receiver

The goal is to confirm the two most important aspects of the PEP, a faster completion time and little to no interference with the interactive traffic. To achieve this we will send a 32 megabyte file from the sender to the receiver, while simultaneously having an interactive UDP traffic sending 100 byte packets at 20 packets per second. The interactive data will simulate Skype traffic using a program called ULTRA_PING[22] which offers end to

end latency measurements.

Additionally we will compare BFIFO to FQ-CoDel, BFIFO being the worst case scenario for the interactive traffic and FQ CoDel the optimal choice for our scenario.

5.3.1 Configuration

The testbed is configured to simulate a base case where the PEP would be beneficial. The connection between the sender and router is configured to have a bandwidth of 100Mbit/s with 100ms delay. This simulates a stable wired connection and will not change throughout the experiments.

The connection between the router and the receiver is initially configured with a bandwidth of 15Mbit/s and 20ms delay, this configuration simulates the unstable wireless domain. The 15Mbit/s represents the low bandwidth phase and will change to 70Mbit/s to simulate the high bandwidth phase.

Delay is configured on interfaces facing from Receiver to Sender, while bandwidth is configured from the Sender to Receiver facing interfaces.

Sender

The sender configures its interface with 100mbit rate using HTB.

```
$ sudo tc qdisc add dev 10Ge root handle 1: htb default 11;  
$ sudo tc class add dev 10Ge parent 1: classid 11 htb rate 100mbit;
```

Router (PEP)

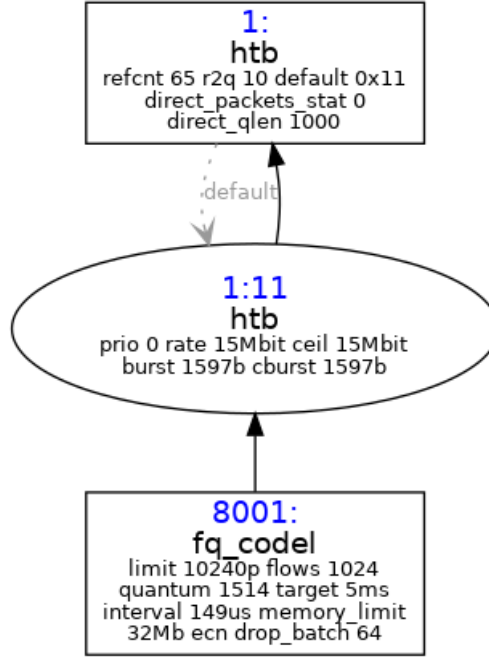
For the Sender facing interface we simply apply 100ms delay:

```
$ sudo tc qdisc add dev 10Gc root handle 2: netem delay 100ms;
```

The sender has multiple configurations for the receiver facing interface based on if we intend to use FQ-CoDel, shown in Figure 5.3, or BFIFO, shown in Figure 5.4.

The BFIFO queue is set to have a limit of the end to end bandwidth delay product (BDP).

- Bandwidth (initial): $B_1 = 15$ Mbit/s



```
$ sudo tc qdisc add dev 10Gf root handle 1: htb default 11;
$ sudo tc class add dev 10Gf parent 1: classid 11 htb rate 15mbit;
$ sudo tc qdisc add dev 10Gf parent 1:11 fq_codel interval 150;
```

Figure 5.3: FQ-CoDel traffic control visualization

- Delay: $D = 120$ ms
- Bandwidth (after increase): $B_2 = 70$ Mbit/s

The Bandwidth-Delay Product (BDP) is calculated as:

$$BDP = Bandwidth \times Delay \quad (5.1)$$

For the initial bandwidth:

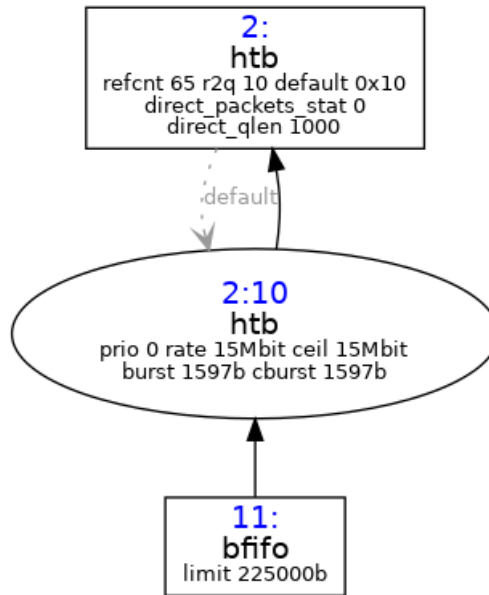
$$BDP_1 = 15 \times 10^6 \text{ bits/s} \times 120 \times 10^{-3} \text{ s} = 225 \times 10^3 \text{ bytes} \quad (5.2)$$

Hence, the initial BFIFO limit is set to 225,000 bytes.

After the increase in bandwidth:

$$BDP_2 = 70 \times 10^6 \text{ bits/s} \times 120 \times 10^{-3} \text{ s} = 1,050 \times 10^3 \text{ bytes} \quad (5.3)$$

Therefore, the new step BFIFO limit becomes 1,050,000 bytes.



```
$ sudo tc qdisc add dev 10Gf root handle 2: htb default 10;
$ sudo tc class add dev 10Gf parent 2: classid 10 htb rate 15mbit;
$ sudo tc qdisc add dev 10Gf parent 2:10 handle 11: bfifo limit 225000;
```

Figure 5.4: BFIFO traffic control visualization

Receiver

The only configuration for the receiver is the 20ms delay on its router facing interface.

```
$ sudo tc qdisc add dev 10Ge root handle 2: netem delay 20ms;
```

5.3.2 Experiment Procedure

The file transfer is simulated by a two custom C programs. We use a command line argument to specify if the Sender uses the `pep_connect` or `connect` function. The reason for creating custom C programs is that we need to use the TLV library for connecting to the PEP. The sender will use the default socket API, sending 1500 bytes at a time. The receiver is responsible for receiving the data and keeping track of the time it took. The receiving server can handle multiple file transfers at the same time using threads.

The experiment consists of three steps:

- **Step 1:** In the first step we start ULTRA_PING from the sender to the receiver and wait 5 seconds to get a baseline of the interactive flow without any disruptions.
- **Step 2:** The second step consists of starting the file transfer. The file transfer can be configured to either be end to end or utilize the PEP.
- **Step 3:** After 13 seconds we change the bandwidth between the router and the receiver from 15Mbit/s to 70Mbit/s, simulating fluctuating bandwidth.

The experiments will be run as shell scripts from the controller host, using `ssh` to execute commands on sender and router. The reason for using scripts is to keep the experiments as consistent as possible.

5.3.3 Results & Analysis

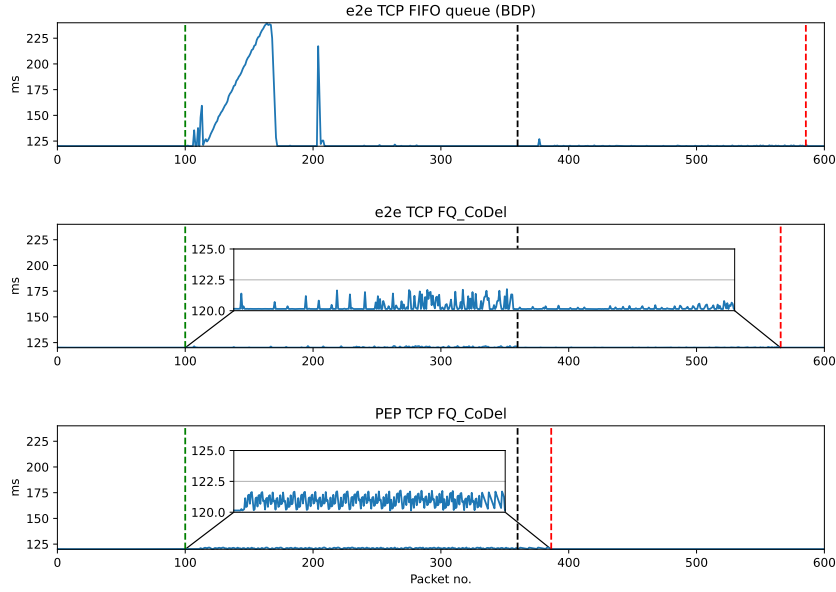


Figure 5.5: File transfer impact on interactive traffic

Figure 5.5 shows three interactive flows, each competing with a file transfer. The first is a normal end to end transfer using the FIFO (bfifo) queue at the router. The second is an end to end transfer using FQ-CODEL, and the last is the same transfer with the PEP.

The lines represent important points in time for each flow, the green line is the start of the file transfer which starts after 5 seconds. The black line shows the the step function which changes the bandwidth from 15Mbit/s to 70Mbit/s. The last red line represents the file transfer completion time.

An important aspect of Figure 5.5 is the big impact that the BFIFO queue has on the interactive traffic. The interactive traffic spikes up to 240ms, which is the normal behavior with big network buffers and the reason buffers normally are kept small. When using FQ-CoDel the interactive flow is barely disrupted with and without PEP.

Based on the last two graphs of Figure 5.5, we can see that the PEP does not disrupt the interactive traffic more than the default end to end transfer, even though the PEP has large buffers which reduce the completion time by 38.54%. See Table 5.1 for all completion times.

Flow	FCT
E2E BFIFO	25.30s
E2E FQ CoDel	23.29s
PEP FQ CoDel	14.31s

Table 5.1: Table showing the Flow Completion Times (FCT) for each file transfer.

Another important aspect which is not visible on the graph is the completion time of the Sender to PEP transfer. Since we split the connection into two transfers, when we use a PEP, we get two completion times. The connection between the Sender and PEP has a much higher bandwidth of 100Mbit/s in comparison with the total end to end bandwidth of 15Mbit/s, which leads to a completion time of 1.5s. This leads to both a better bandwidth utilization as discussed in Chapter 3, and frees the potentially contested bandwidth faster.

Figure 5.6 shows a closeup of the end to end and PEP transfers using FQ-CoDel. Although there are more fluctuations when using the PEP, the delay variation stays inside the same range of 120.0ms to approximately 121.5ms. Its important to note that the PEP transfer finishes earlier, which means it impacts the interactive traffic for a shorter amount of time than the end to end transfer.

Table 5.2 shows the average, median and max delay for both the PEP and end to end transfer. The statistics are taken over the time period of

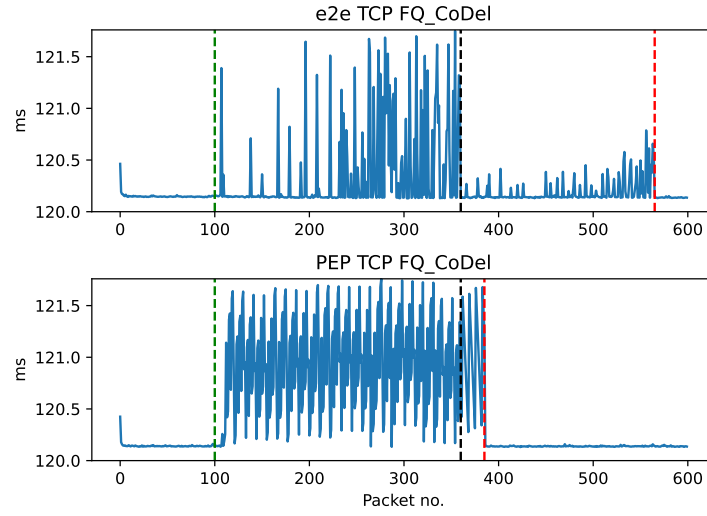


Figure 5.6: Close up of End to End compared to PEP

	Average	Median	Max
E2E	120.29ms	120.14ms	121.74ms
PEP	120.61ms	120.40ms	121.75ms

Table 5.2: Delay statistics (Calculated based on longest completion time)

the longest flow to include the benefit of finishing early for the PEP. This comes to a total of 0,322ms difference on the average and 0,257ms on median.

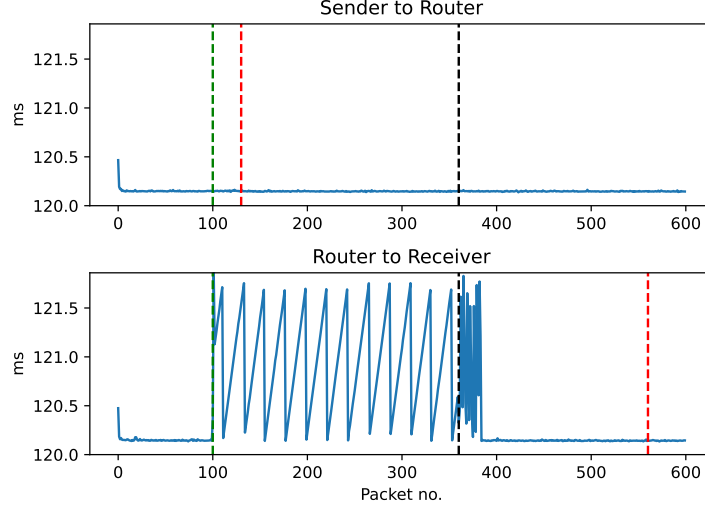


Figure 5.7: Sender to Router and Router to Receiver effect on the interactive traffic.

To demonstrate where the delay happens, we can look at two individual flows: one from the Sender to the Router and one from the Router to the receiver, each paired with the end to end (Sender to Receiver) UDP interactive traffic. Figure 5.7 shows the resulting graphs; as we can see, there is almost no disruption on the Sender to Router transfer. Most delay comes from the bottleneck link between the Router and Receiver.

5.4 Using real 5G traces

The next experiment will focus on the PEPs behavior in the context of realistic 5G mmWave traces. The focus will be on the throughput over a simulated 5G network with and without the PEP. Our experiments will follow up on earlier experiments by Witestlab, where real 5G 60GHz WLAN traces were collected from a testbed [25] [11]. The Witestlab experiments include four different 5G scenarios which we will first validate and then use to evaluate the performance of the PEP.

5.4.1 Traces

The traces were collected by using two laptops with 60GHz wireless Qualcomm cards, using iperf3 to have a loaded link [25]. The traces were captured

over 120 seconds under four different scenarios. In [25], they are described as follows:

- **Static Link:** In the first scenario, the laptops are stationary and nothing obstructs their path.
- **Short Blockages:** Same as the static link scenario, however at a 15 second interval a human obstructs the path between the laptops.
- **Long Blockages:** This scenario is similar to the short blockages, however in this scenario the human stopped on the path for four seconds each interval.
- **Mobility and Blockages:** In the last scenario, the client laptop moved in a circle around the room, while a human obstructed the signal path at a 15 second interval.

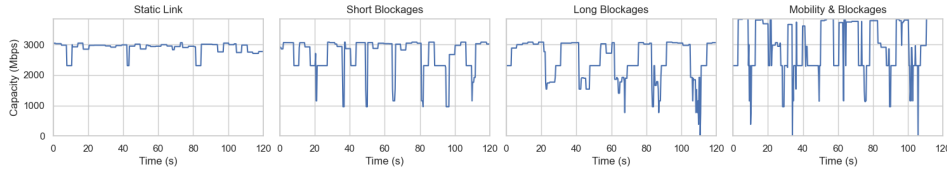


Figure 5.8: mmWave capacity fluctuation for four different 5G scenarios (Data from [25])

Figure 5.8 visualizes the fluctuating bandwidth. The traces have a bandwidth of approximately 1Gbit/s to 4Gbit/s based on the scenario, even reaching zero with the mobility and blockages scenario. The raw data of the traces are contained in .csv files which consist of a time difference and the capacity of the link in megabits.

5.4.2 Topology

For this experiment we will use a topology similar to the previous experiment. We will continue with our sender, router and receiver topology, however, the traffic control configuration will change. Because of the high 5G bandwidth we will adjust the sender to router link to have the maximum our testbed can offer, which is a 10Gbit/s bandwidth.

The link between the sender and the router will be configured based on the scenario, but the default will be a 3Gbit/s bandwidth.

5.4.3 Configuration

Since we are now dealing with high gigabit bandwidths, we need to configure and tune the hosts to be able to fully utilize these bandwidths. Most importantly, the default network read and write buffers need to be adjusted.

The core net buffers `net.core.rmem_max` and `net.core.wmem_max` will be adjusted to approximately 256 megabytes. Also `net.ipv4.tcp_rmem` and `net.ipv4.tcp_wmem` will be set to the same 256 megabytes size. This will assure that neither the sender or receiver is limited by the socket buffers.

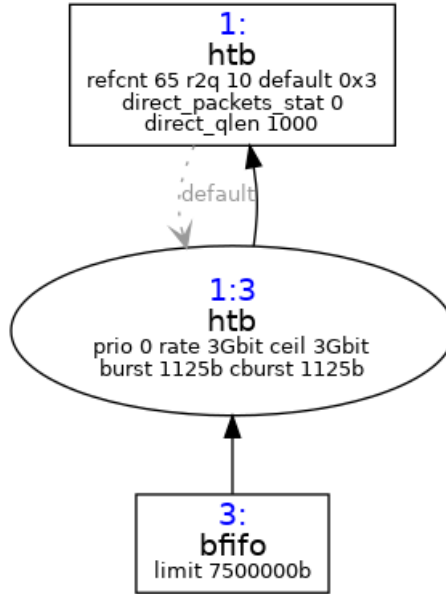
5.4.4 Validation

First we will try to reproduce the results from the original Witestlab experiment, to see how our testbed will affect the results. Following each step of the original experiment, we configure the link between the router and the receiver to initially have a 3Gbit/s bandwidth and no configured delay. Additionally, it will have a BFIFO queue with a max capacity of 7,5Mb which will create a delay of 20ms if the queue is fully filled. This configuration is visualized in Figure 5.9.

To simulate the fluctuating bandwidth, a shell script at the router will configure the `tc htb rate` at a given interval from a `.csv` file. The sender will use `iperf3` to create ten TCP flows which will fully utilize the available bandwidth. To measure the latency impact of the 5G traces we will use the `ss` command to gather the RTT's of each flow. In our validation we will only run our experiments using Cubic as it is the system default.

Figure 5.10 shows the latency results from the Witestlab. We will focus on their TCP Cubic results as it is the congestion controller we use for our experiments. When using Cubic, Witestlab experience delays up to 150ms in the worst cases for long blockage.

The results from our latency measurement experiment show a clear impact on the latency when the bandwidth fluctuates. With a static link the delay reaches approximately 20ms, which is the result of the 7,5Mb BFIFO queue at the router. The steady 20ms delay for the static link is the same result Witestlab observes. However, in our experiments we reach a maximum of approximately 50ms delay with short and long blockages, where Witestlab reaches 150ms.



```

$ tc qdisc del dev 10Gf root;
$ tc qdisc replace dev 10Gf root handle 1: htb default 3;
$ tc class add dev 10Gf parent 1: classid 1:3 htb rate 3gbit;
$ tc qdisc add dev 10Gf parent 1:3 handle 3: bfifo limit 7500000;

```

Figure 5.9: BFIFO traffic control visualization for 5G traces



Figure 5.10: Results from the Witestlab experiment [25]

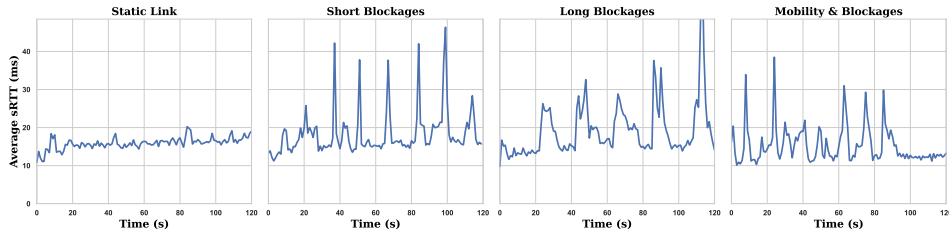


Figure 5.11: Our results for the delay caused by fluctuating bandwidth

Figure 5.11 shows the results of the latency measurements. Comparing our results to the Witestlab experiment (Figure 5.10) we can see that the shapes of the graphs are similar. However the magnitude of the latency spikes are quite different between our and their experiments. The difference is most likely due to our hardware or a more modern Linux kernel.

Based on these results we can conclude that the 5G traces reflect the expected behavior: creating immense delays in the case of a big BFIFO buffer. Although we do not observe the same amount of delay, we still observe the same general latency behaviour of the 5G traces.

5.4.5 Experiment Configuration

To evaluate the PEP with the 5G traces we will change the topology to include a 100ms delay between the sender and the router. The link between the router and the receiver will get a 1ms delay for a more realistic scenario. The experiment procedure will be to send as much data as possible from the sender to the receiver. We will use throughput as the main measurement for the performance of each scenario. The interactive traffic should not, or hardly be, impacted by the other transfer(s). It will be represented by Ultra ping.

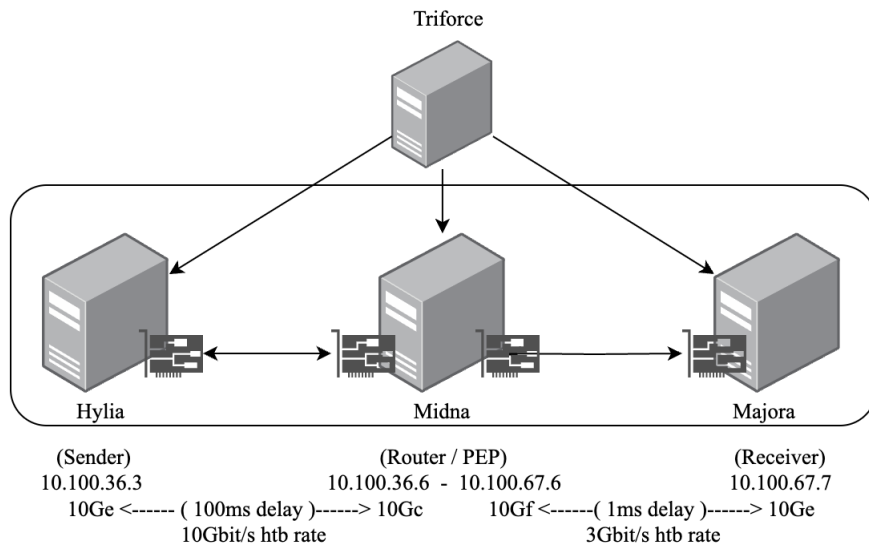


Figure 5.12: Testbed configuration for 5G traces

End-to-end Configuration

The end-to-end scenario has two cases, once with FQ-CoDel and once with BFIFO. FQ-CoDel will reduce the amount of impact on the interactive traffic, while BFIFO will allow the sender to buffer a lot of data at the router. The BFIFO buffer will be configured with a queue size of the BDP based on the static scenario. The total end to end RTT is 101ms and the bottleneck bandwidth is 3gbit/s. Based on the calculation from Figure 5.13 the BDP is 37,875,000 bytes. This will allow for a lot of data to be buffered at the router, which will negatively impact the interactive traffic, but increase the throughput.

$$\begin{aligned}\text{BDP (in Bytes)} &= \frac{\text{Bandwidth (in bits/s)} \times \text{RTT (in s)}}{8} \\ &= \frac{3 \times 10^9 \times 0.101}{8} \\ &= 37,875,000 \text{ Bytes}\end{aligned}\tag{5.4}$$

Figure 5.13: Calculation of the Bandwidth-Delay Product (BDP) in Bytes

The 100ms delay will be configured on the Sender's interface towards the router, while the 1ms delay will be configured on the receivers interface.

PEP Configuration

For the PEP case we will change the configuration to assert that no unwanted delay is added for the interactive traffic from the sender. We will achieve this by adding FQ-CoDel at the sender's interface and configuring the 100ms delay on the router's ingress interface for traffic from the sender (in the BFIFO case we configured the 100ms on the sender's outgoing interfaces). Because the PEP essentially will be a split connection at the router we must be sure that interactive traffic is not already delayed before it reaches the PEP.

Especially important for the PEP are the socket buffers at the router, which are the buffers that the PEP will use. These buffers would normally be ignored by packets simply passing through the router. Therefore they will not affect any interactive traffic, as NIC or scheduler buffers would. The socket buffers on the router will be configured to be 128 megabytes, which will be enough to temporary buffer packets for low bandwidth phases. It is

noteworthy that a socket has two buffers, one for reading and one for writing. Additionally a connection has two sockets on the PEP, which gives us a total of four buffers.

5.4.6 Experiment Steps

For this experiment we will have two parallel flows, one for the data and one for the interactive traffic.

The experiment consists of three steps:

- **Step 1:** The router starts the 5G fluctuations trace simulations.
- **Step 2:** Ultra ping starts sending interactive traffic from sender to receiver.
- **Step 3:** The sender will start transferring data to the receiver.

Congestion Controller's affect on throughput

While running a throughput test from sender to receiver with a BDP sized BFIFO queue at the router, we observe surprising results of throughput. We expected the throughput, with our configuration, to be about 3Gbit/s: fully utilizing the bandwidth. However the results hover around 2.1Gbit/s to 2.5Gbit/s. To understand this we can look at the congestion window of a cubic transfer.

The first interesting aspect of Figure 5.14 is the slow cubic curve. It takes a long time to fully utilize the bandwidth. Additionally, looking at the relation to the BDP, we can see that the transfer was lucky in the first increase phase, starting just under the BDP, and unlucky the second where it starts way below the BDP, resulting in a rather low throughput.

Looking at the double back-off after 70 seconds in Figure 5.14 we can see that the next congestion avoidance phase starts low and does not reach full throughput again before the experiment ended. This brings us to the conclusion that Cubic does not operate effectively in our scenarios with big bandwidths and large buffers. Because of these results a single Cubic transfer is not representative of the achievable throughput as it will either be lucky and get a big throughput or unlucky with a low throughput.

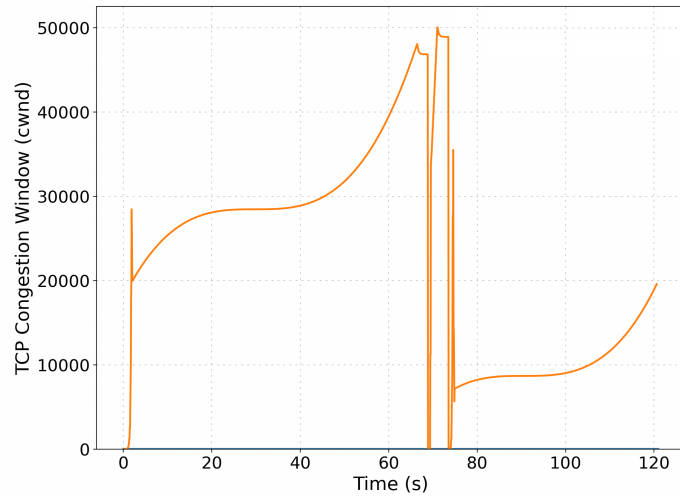


Figure 5.14: Cubic's congestion window

Looking at Reno, we see the same kind of behavior, where the big BFIFO queue and high bandwidth results in low throughput because of the congestion avoidance phase. In Figure 5.15 we can see an unlucky congestion avoidance start which leads to low throughput.

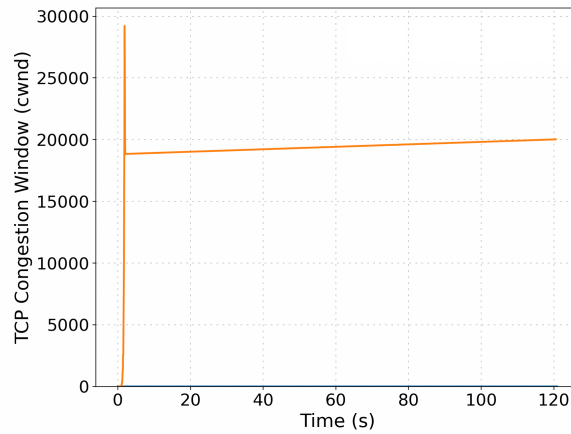


Figure 5.15: Reno's congestion window

Because of the unstable nature of Cubic in our experiments we will run each experiment ten times to avoid a lucky or unlucky throughput result.

5.4.7 Results

The results will focus on the differences in throughput based on the scenario. Importantly, how the choice of buffer, scheduler and 5G scenario affect the fairness, impact on interactive traffic and achievable throughput.

Static link

The static link scenario will represent a baseline of performance with little to no fluctuations. Based on this baseline we can compare the effect of the various types of fluctuations. The static link configuration has an average throughput of 2905.61 Mbit/s and a standard deviation of 171.74 Mbit/s.

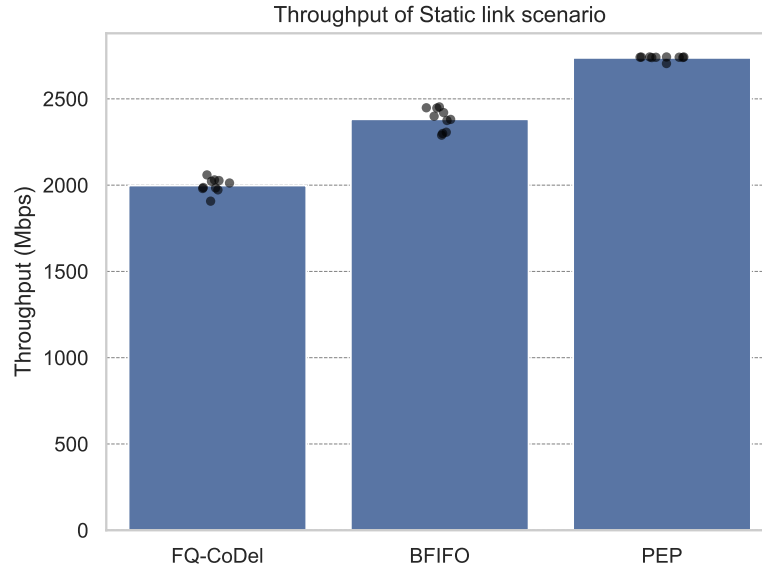


Figure 5.16: Bar plot of throughput for each scenario with a static link configuration

Based on the small standard deviation and small fluctuations, cubic will not be expected to suffer under this configuration. Looking at the throughput in Figure 5.16, we can see that there is only a small difference in throughput for each configuration. FQ-CoDel has the lowest overall throughput which is the result of lower buffer sizes as to not disrupt the interactive traffic. In contrast the BFIFO configuration, with a BDP sized buffer, has higher throughput at the cost of disrupting the traffic.

The PEP has the highest throughput of 2737.83 Mbit/s, almost reaching the average throughput of the scenario. The PEPs throughput highlights the benefits of both large buffers and a split connection.

	mean	std
FQ-CoDel	1998.12 Mbit/s	42.10 Mbit/s
BFIFO	2382.00 Mbit/s	63.76 Mbit/s
PEP	2737.83 Mbit/s	11.67 Mbit/s

Table 5.3: Statistical analysis of throughput data for a static link configuration

Looking at the statics in Table 5.3 we can see a rather low standard deviation of the throughput for each scenario. BFIFO has the largest deviation, most likely due to the aforementioned Cubic behavior with large buffers and high delay. The PEP has the lowest deviation of 11.67 Mbit/s, another benefit of a split connection.

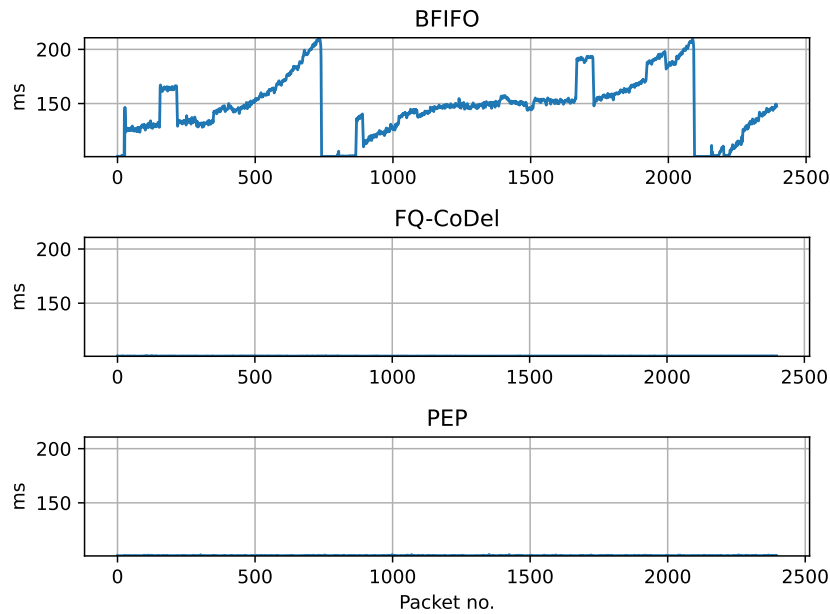


Figure 5.17: Impact on interactive traffic for the static link scenario

Figure 5.17 shows the impact on the interactive traffic for each configuration. Looking at the BFIFO configuration we get a big impact on the

interactive traffic because of the large buffers. This shows the cost of the higher throughput that BFIFO achieves over FQ-CoDel. While the base RTT of our configuration is 101ms, the BFIFO reaches peaks above 200ms and a average of 146.83ms with a 26.59ms standard deviation.

The impact of the FQ-CoDel and PEP configuration is barely noticeable, compared to the BFIFO configuration. Figure 5.18 shows a close up of both FQ-CoDel and PEP; as we can see, there is a tiny difference with the PEP having more spikes than FQ-CoDel.

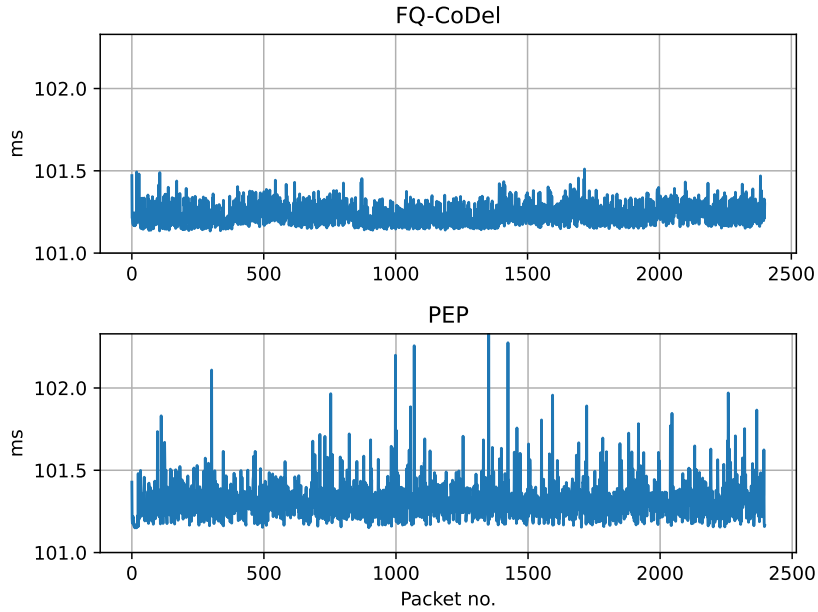


Figure 5.18: Closeup of the PEP and FQ-CoDel impact on interactive traffic for the static link scenario

Although the difference between FQ-CoDel and PEP may visually appear large, it is insignificant: there is only a difference of 0.078ms for the mean, and 0.047ms difference for standard deviation. Table 5.4 shows the mean and standard deviation for all three configurations. This highlights one of the main benefits of the PEP. We achieve a higher throughput than BFIFO while barely impacting the traffic more than FQ-CoDel would.

	Mean Latency (ms)	Standard Deviation (ms)
BFIFO	146.835	26.598
FQ-CoDEL	101.236	0.062
PEP	101.314	0.109

Table 5.4: Impact on latency statistics

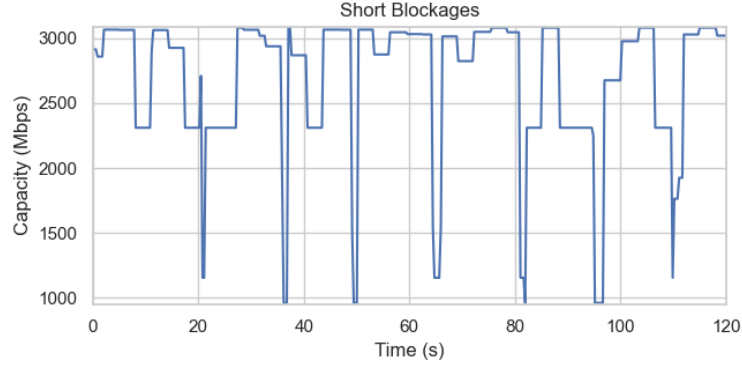


Figure 5.19: Bandwidth fluctuations with Short blockage scenario

Short blockages

The short blockage is an important scenario to evaluate as it will show the impact that small and short fluctuations will have on the throughput, especially for FQ-CoDel. The average throughput of the short blockage configuration is 2726.27 Mbit/s with a 517.02 Mbit/s standard deviation.

From Figure 5.20 we can see an overall decrease of throughput, which is expected based on the average throughput of the scenario. However, FQ-CoDel has an especially big decrease in throughput from a 1998.12 Mbit/s average throughput in static link down to 1017.17 Mbit/s in the short blockage scenario. In this scenario the benefit of big buffers becomes especially apparent, the small buffer sizes of FQ-CoDel with fluctuations forces FQ-CoDel to more frequently drop packets, which is devastating for the congestion controller.

Table 5.5 compares the throughput of each configuration, and shows the drastic drop in throughput for FQ-CoDel. Additionally, FQ-CoDel has a higher standard deviation compared to the static link scenario. Both BFIFO and the PEP are able to buffer more packets, which makes them more tolerant to fluctuations.

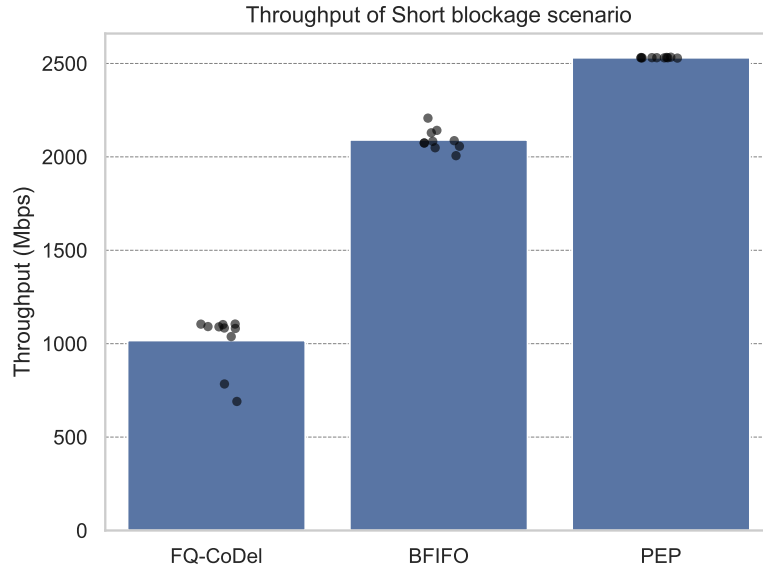


Figure 5.20: Bar plot of throughput for each scenario with a short blockage configuration

	mean	std
FQ-CoDel	1017.17 Mbit/s	150.29 Mbit/s
BFIFO	2090.83 Mbit/s	56.05 Mbit/s
PEP	2530.90 Mbit/s	1.33 Mbit/s

Table 5.5: Statistics of the throughput for the short blockage scenario.

The impact on interactive traffic does not change much for FQ-CoDel or PEP, however the BFIFO configuration creates big spikes with short blockages. Figure 5.21 visualizes the impact of BFIFO on the interactive traffic compared to the static link results. With the short blockage scenario the latency of the interactive traffic reaches up to 400ms, this is approximately double the latency peak that the static link scenario reached.

	Mean Latency (ms)	Standard Deviation (ms)
Short blockage	152.79	54.99
Static Link	146.83	26.59

Table 5.6: Latency statistics for BFIFO with short blockage and static link.

Table 5.6 shows that the overall impact is not as big as the graph may

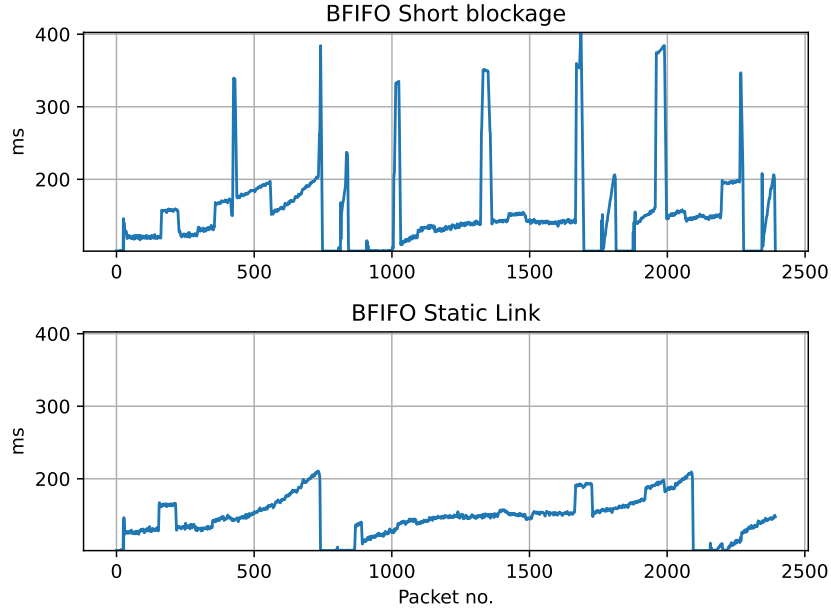


Figure 5.21: BFIFO impact on interactive traffic with short blockage and static link.

indicate. The average difference is only 5.96ms, with a 28.4ms difference for standard deviation. However, even if only rarely, a delay increase from 101ms to 400ms is still a disruption to any interactive experience.

	Mean Latency (ms)	Standard Deviation (ms)
PEP	101.30	0.09
Static Link	101.18	0.06

Table 5.7: Latency statistics for FQ-CoDel and PEP with short blockage

Long blockages

The long blockage scenario is special as it has the lowest average throughput of 2519.73 Mbit/s and highest standard deviation of 621.16 Mbit/s. The blockage lasting for a longer time reduces how fast the bandwidth fluctuates.

Looking at Figure 5.22 we can see a positive effect with the FQ-CoDel configuration compared to Figure 5.20. For FQ-CoDel the rapid fluctuations are the worst case scenarios, because it is unable to react fast enough or buffer

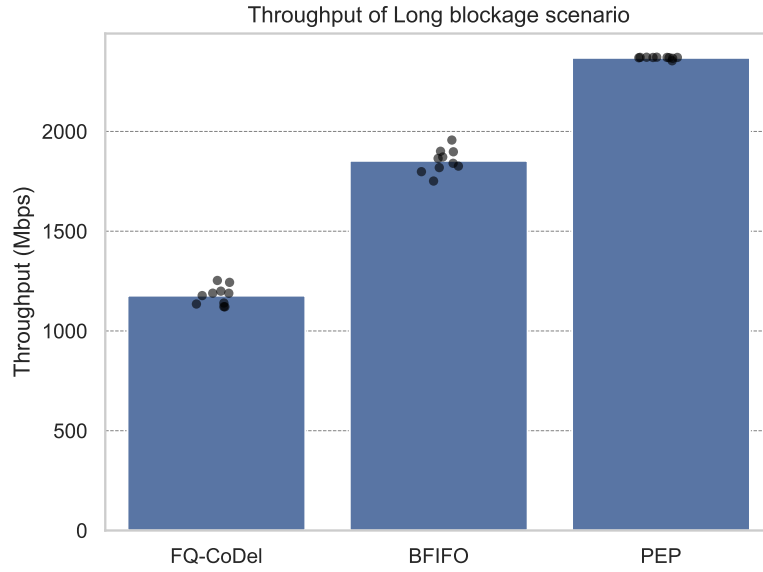


Figure 5.22: Bar plot of throughput for each scenario with a long blockage configuration

enough packets as to not drop them. In table 5.8 we can see that the FQ-Codel throughput increased by 159.80 Mbit/s while the standard deviation decreased by 102.74 Mbit/s compared to the short blockage scenario.

	mean	std
FQ-CoDel	1176.97 Mbit/s	47.55 Mbit/s
BFIFO	1852.70 Mbit/s	58.61 Mbit/s
PEP	2368.77 Mbit/s	5.36 Mbit/s

Table 5.8: Statistics of the throughput for the long blockage scenario.

	Mean Latency (ms)	Standard Deviation (ms)
PEP	101.33	0.11
FQ-CoDel	101.21	0.08
BFIFO	144.56	60.55

Table 5.9: Latency statistics for long blockage scenario.

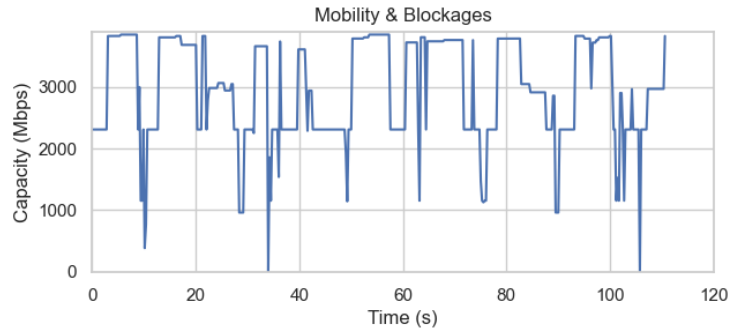


Figure 5.23: Bandwidth fluctuations with mobility and blockage scenario

Mobility and Blockages

The mobility and blockage scenario will be the most interesting scenario for our experiment, as it captures the volatile nature of 5G mmWave networks. Because of high fluctuations we get an average throughput of 2962.61 Mbit/s which is the highest average throughput of all scenarios. This comes at the cost of the highest standard deviation of almost 1Gbit/s, more exactly 826.45 Mbit/s.

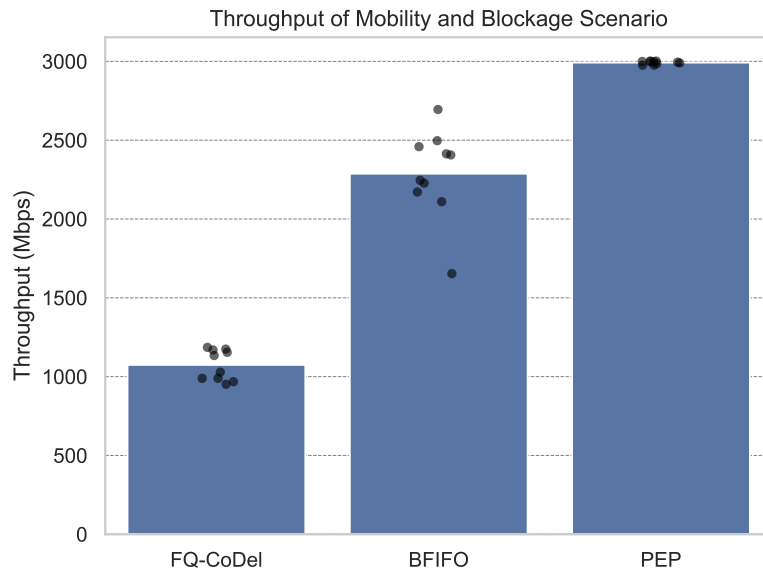


Figure 5.24: Bar plot of throughput for each scenario with the mobility and blockage scenario

Figure 5.24 shows the throughput of each configuration, most notably the

PEP has a big increase from the previous scenario. The PEP has an average throughput of 2992.35 Mbit/s which is over the average of the scenario. This result clearly shows the benefit of the PEP, especially in a highly fluctuating environment as in this scenario. The PEP is able to utilize the high bandwidth phases by buffering the most data and reacting faster than any other configuration.

The BFIFO configuration also has an increase of 435.05 Mbit/s from the previous scenario, in contrast to FQ-CoDel which had a decrease in throughput compared to the previous scenario. This highlights the importance of larger buffers in highly fluctuating environments. However, the BFIFO configuration suffers from a much higher standard deviation of 283.13 Mbit/s and an approximate difference of 1Gbit/s between the Min and Max throughput results.

	mean	std
FQ-CoDel	1074.42 Mbit/s	96.78 Mbit/s
BFIFO	2287.75 Mbit/s	283.13 Mbit/s
PEP	2992.35 Mbit/s	10.07 Mbit/s

Table 5.10: Statistics of the throughput for the mobility and blockage scenario.

Looking at the impact on latency we see a similar result to the static link and short blockage scenarios. BFIFO has a big impact on the interactive traffic, especially in fluctuating scenarios, while both the PEP and FQ-CoDel show little to no impact.

Figure 5.26 shows a closeup of the impact FQ-CoDel has in comparison to the PEP. Interestingly both have a big spike at the start, most likely due to slow start, but manage to keep the impact minimal for the rest of the experiment. The PEP seems to be more volatile than the FQ-CoDel, which most likely is due to the PEP reacting to the fluctuations and thereby achieving the high bandwidth. Considering the PEP almost has double the throughput FQ-CoDel has, the latency impact is negligible.

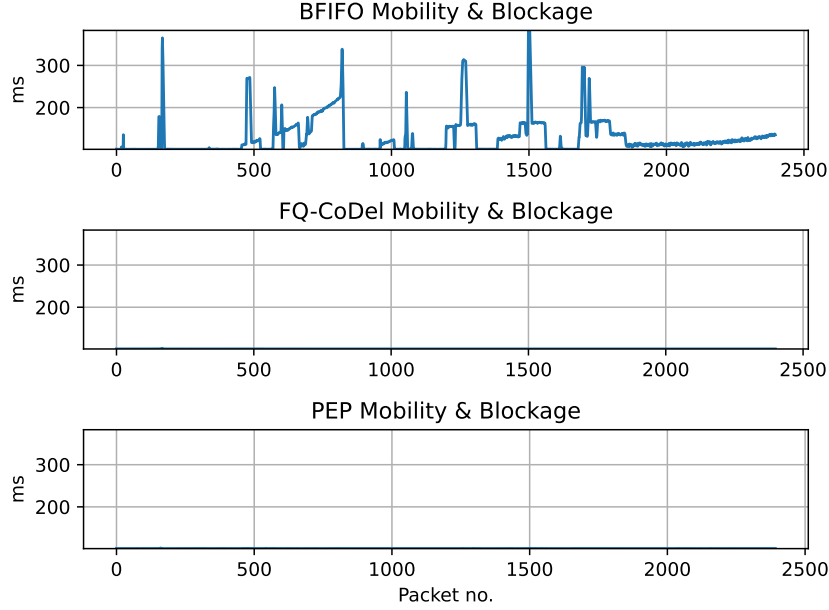


Figure 5.25: Impact on interactive traffic for each configuration in the mobility and blockage scenario.

5.4.8 Evaluation of 5G traces experiment

After evaluating the PEP with realistic 5G mmWave traces we can clearly see the traditional dilemma of buffer sizes versus impact on interactive traffic, especially with fluctuating bandwidth. When configuring BFIFO with a BDP sized queue at the router, we see a clear performance increase in the throughput compared to FQ-CoDel with small buffer sizes.

Although BFIFO achieves higher throughput than FQ-CoDel, the interactive latency traffic suffers greatly under the large buffers. As we can see in Figure 5.27, BFIFO creates immense amounts of delay, reaching peaks of 400ms, for all scenarios.

The PEP achieves both high throughput and low impact on interactive traffic by combining large buffers with FQ-CoDel and a split connection. The socket buffers do not disrupt traffic which passes through and FQ-CoDel assures a fair and low impact delay. Especially in the mobility and blockage scenario, the PEP is able to almost fully utilize the available bandwidth.

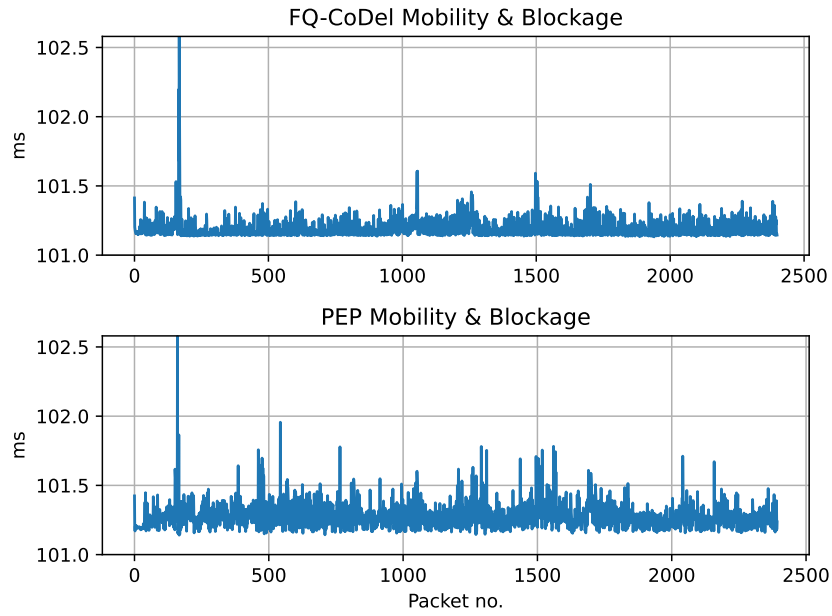


Figure 5.26: Closeup of the PEP and FQ-CoDel impact on interactive traffic for the mobility and blockage scenario

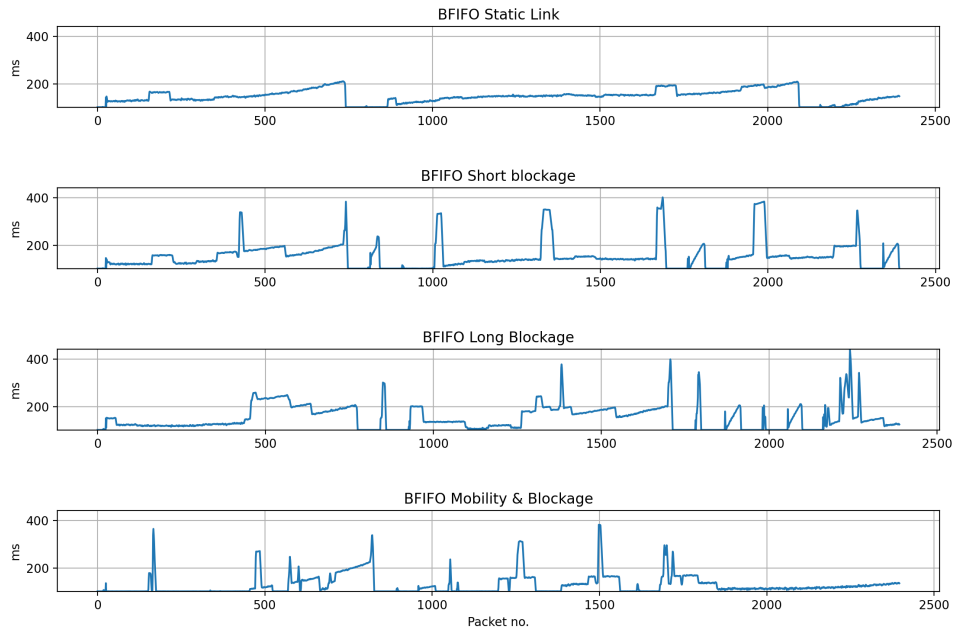


Figure 5.27: BFIFO impact on latency.

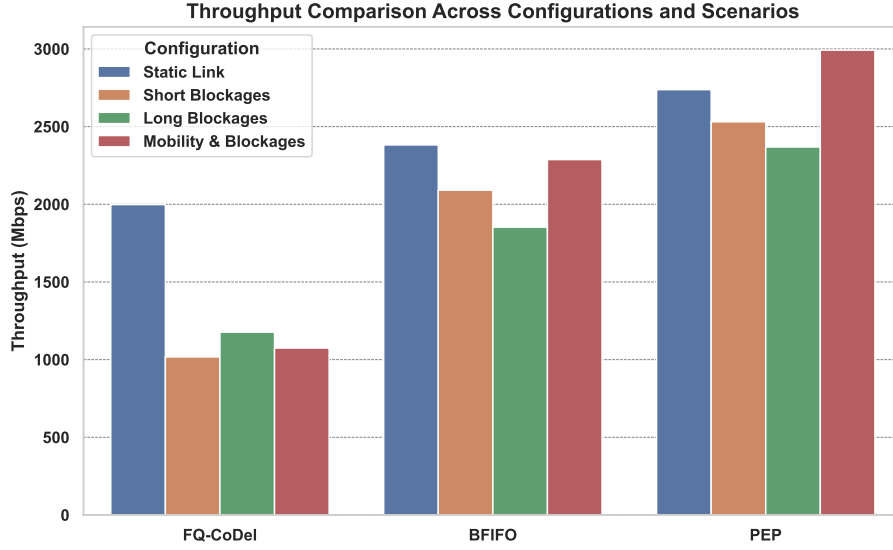


Figure 5.28: Average throughput for each scenario by configuration.

5.5 Socket buffer size effect on throughput

Continuing on the Mobility and Blockage scenario from the last section, we will evaluate how the socket buffer sizes on the PEP affect the performances. For this experiment we will run PEP configuration with ten different socket buffer sizes. The evaluation will consist of ten steps, ranging from the default Linux (debian) socket buffers up to 128 megabytes.

The Linux default socket buffer sizes start at 2 megabyte for net core read and write buffers. The TCP read buffers are at 6 megabyte, and the write buffers at 4 megabyte. The default buffers will be increased at a 1/10 interval until they reach 128 megabyte. Each step, we will run one flow using the PEP for the entire 120 seconds. For a consistent result we will run each step ten times and take the average as the final result. The goal is to evaluate and confirm the buffers effect on the PEP, compared to the FQ-CoDel and BFIFO configuration.

As we can see from Figure 5.29, the default socket buffer sizes perform the worst, improving with each step. From the default to step 7 the throughput increases with bigger socket buffer sizes. This highlights the importance of the buffer sizes for the PEP. After step 7, approximately 85.40 MB buffer size, the throughput reaches its limit based on the scenario. The limit is reached by having enough data buffered to compensate for drop and rise in

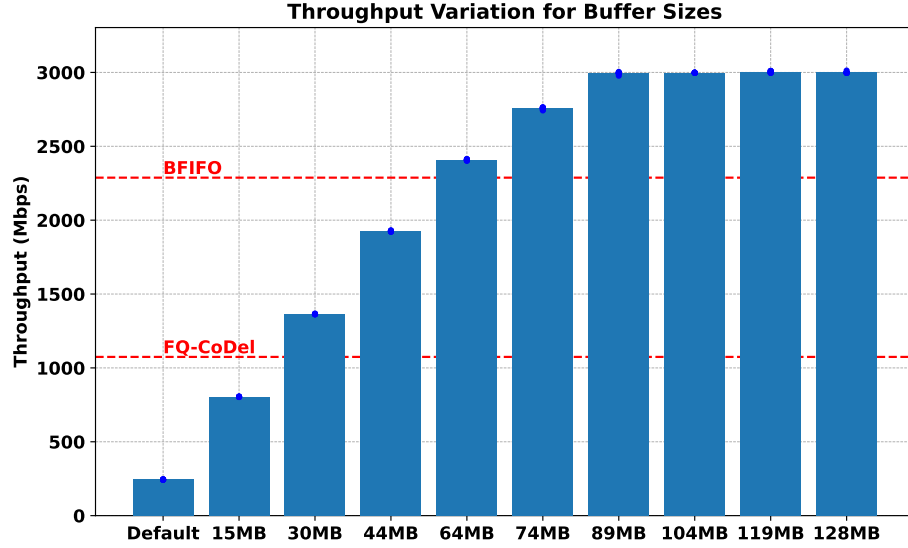


Figure 5.29: Throughput for the PEP based on socket buffer sizes in ten steps.

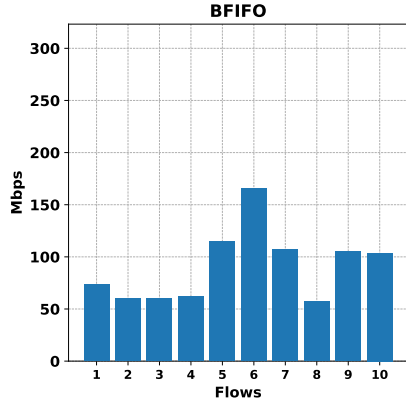
bandwidth.

5.6 Multiple flows

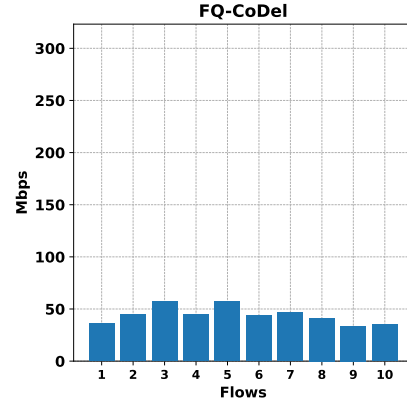
The last section of the Evaluation chapter will focus on the PEP performance with multiple flows. Especially throughput and fairness are important aspects to explore. The mobility and blockage scenario from the 5G mmWave section will be used to simulate our fluctuating bandwidth.

Figure 5.30 shows the ten flows for each of the three configurations. The main takeaways are the loss in total throughput for BFIFO and FQ-CoDel, where having multiple flows compete with the heavy fluctuations leads to less bandwidth utilization. per flow. The PEP only has minimal throughput loss, even with multiple flows.

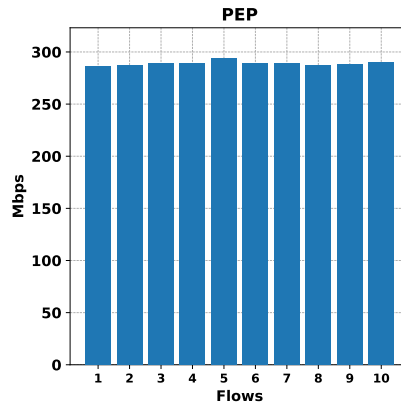
Another interesting aspect is the flow fairness: as we can see in Figure 5.30, with BFIFO there is a clear unfairness between the ten flows. The Jain's fairness index under each graph indicates the fairness, an index of 1 is considered very fair, while the number diminishes as fairness degrades. One of the beneficial side-effects of FQ-CoDel is its inherent fairness, as we can



(a) Throughput for 10 flows with BFIFO (Jain's fairness index 0.8827)



(b) Throughput for 10 flows with FQ-CoDel (Jain's fairness index 0.9694)



(c) Throughput for 10 flows with PEP (Jain's fairness index 1.0000)

Figure 5.30: The throughput of ten parallel flows for each configuration.

	mean	std
FQ-CoDel	44.12 Mbit/s	7.84 Mbit/s
BFIFO	90.89 Mbit/s	33.13 Mbit/s
PEP	289.10 Mbit/s	1.89 Mbit/s

Table 5.11: Average throughput of a single flow

see in Figure 5.30 (b). Because the PEP also uses FQ-CoDel, we get the same fairness, but without the loss in throughput.

Table 5.12 and Figure 5.31 show the impact on the interactive traffic

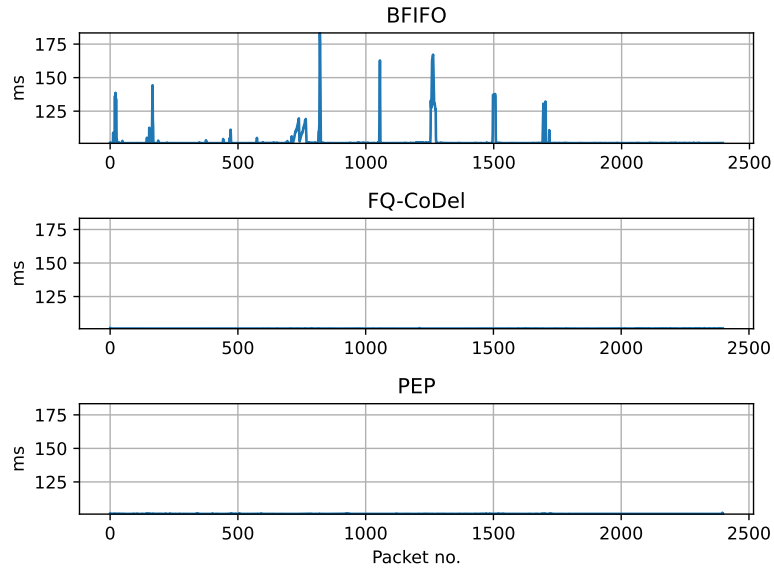


Figure 5.31: Impact on interactive traffic for ten flows with mobility and blockage scenario.

	Mean Latency (ms)	Standard Deviation (ms)
BFIFO	102.37	6.35
FQ-CoDel	101.18	0.04
PEP	101.18	0.05

Table 5.12: Latency statistics for ten flows with the mobility and blockage scenario.

with ten flows. Interestingly, the competition between the ten flows create an overall less destructive impact on the interactive traffic than a single flow. With BFIFO there are still high spikes up to 175ms. While both PEP and FQ-CoDel have the same average latency of 101.18ms, which is only an increase of 0.18 from the set end to end delay of 101ms.

Chapter 6

Conclusion & Future Work

In the final chapter we will discuss the PEP's performance, limitations and challenges. It is especially important to what degree we achieved the PEP's goal of increasing throughput without disrupting interactive traffic. After concluding on performance, we will look at the potential challenges and limitations which could reduce the PEPs usability. Finally, potential future work is discussed.

6.1 Performance

To conclude on the PEP, let us look back at the original research question: *Can we get better performance without a negative influence on latency by adding large buffers inside highly fluctuating networks?*

The goal of the PEP is to increase the throughput of non-interactive transfers by better utilizing the fluctuating bandwidth; in this thesis we used 5G mmWave as an example. Our PEP achieves this by introducing large buffers at the base station, which of course is only viable if we introduce no additional unwanted delay on interactive traffic.

6.1.1 Throughput

In Section 5.4.7: Throughput for a 5G mobility and blockage scenario, based on the evaluation with real 5G mmWave traces, we achieve a very high throughput and bandwidth utilization with the PEP. Especially compared to configurations with BFIFO and FQ-CoDel. We are able to show the immense

benefit large buffers have in these kind of scenarios.

In Section 5.4.7 we can see that the PEP is able to almost fully utilize the average bandwidth of approximately 3Gbit/s, even with the high fluctuations, while FQ-CoDel struggles to even utilize half of the theoretically available bandwidth. Configuring BFIFO with a BDP sized queue, we are able to better utilize the bandwidth, almost doubling the throughput of FQ-CoDel. However, introducing these larger buffers in the network causes a lot of unwanted delay on the interactive traffic.

6.1.2 Impact on interactive traffic

To avoid impact on interactive traffic is a core idea behind the usability of the PEP. Although we achieve high throughput, large buffers in the network cause a large amount of unwanted delay, as we have shown in the BFIFO case. In the initial evaluation we showed that we are able to reduce the impact of BFIFO buffers on the interactive traffic by configuring FQ-CoDel instead, which reduces the impact by giving each flow its own small queue. Although FQ-CoDel solves this problem, it also drastically lowers the throughput.

We were able to show that with our PEP we can utilize both FQ-CoDel and achieve high throughput by using a split connection with large socket buffers. This combination assures that any traffic that willingly utilizes the PEP will have the large buffers, while other traffic only adhere to FQ-CoDel. All traffic from the PEP will also eventually be managed by FQ-CoDel which assures that it will not impact any traffic passing through.

Looking at the impact graphs in Section 5.4.7, especially Figure 5.26 shows the small amount of delay the PEP adds compared to the case with FQ-CoDel only.

6.1.3 Transparency

An important aspect that distinguishes our PEP from most other approaches is the choice of a non-transparent PEP. By allowing the user to actively choose the PEP, we don't break end to end assumptions. In this way, an application can make a conscious decision to use the PEP for a suitable type of traffic only. The main benefit of the transparent PEP is the ease of deployment and needing no modifications on the sender or receiver side.

6.2 Limitations and Challenges

A non-transparent PEP has a lot of benefits, however it also comes with its own set of challenges. The process of choosing and addressing the PEP requires a modification on the sender. PEP selection also becomes a problem, in this implementation it is a hard coded value in the TLV library. Needing to adjust the sender code to utilize the PEP also reduces the adaptability of the PEP.

Experimenting with and simulating high bandwidth 5G networks is a challenge. Configuring Linux with these bandwidths causes weird behavior with Cubic, as discussed in the Evaluation chapter. Additionally, when trying to setup certain scenarios, unexpected behaviors occur based on where delay is configured. These problems make the PEP hard to evaluate under emulated scenarios.

6.3 Future work

For future work we propose a real deployment evaluation, especially using the PEP with a real router and a mobile device. As all experiments in this thesis were conducted using a testbed with simulated bandwidth fluctuations, the results may change in a real environment. This would open up to evaluations with real transfers from website alongside interactive traffic from either video games or video conferencing.

Additionally, a better understanding and evaluation of Cubic's behavior in high bandwidth networks with big buffers would aid to further improve the PEP. Usually networks avoid these large buffers, so more research has to go into evaluating congestion controllers in these scenarios. PEP with FQ-CoDel yielded good results in our evaluations, but alternatives such as Cake or Priority First in First Out (PFIFO) could enhance the PEP in ways we have not discovered yet. A future goal would be to have even less impact on interactive traffic with our PEP than normal FQ-CoDel: the large buffers of our PEP should not only increase the throughput, but could also reduce latency impact further. This thesis focused on fluctuating bandwidths in the context of 5G mmWave networks. Regarding the PEP, other types of fluctuating bandwidth can be beneficial to further research.

6.4 Final Thoughts

From the throughput evaluation we thoroughly investigated the benefit of large buffer sizes with 5G mmWave fluctuating bandwidth. Configuring BFIFO with a BDP sized queue yielded double the throughput of FQ-CoDel, although at the cost of a high throughput standard deviation and a large impact on the interactive traffic. Our PEP is able to combine large buffers with FQ-CoDel to achieve both high and stable throughput and low impact on interactive traffic.

Bibliography

- [1] Operating System Statistics - W3Techs. https://w3techs.com/technologies/overview/operating_system. Accessed: April 26, 2024.
- [2] Saksham Agarwal, Arvind Krishnamurthy, and Rachit Agarwal. Host congestion control. In *Proceedings of the ACM SIGCOMM 2023 Conference*, page 275–287, New York NY USA, September 2023. ACM.
- [3] S. K. Agrawal and Kapil Sharma. 5g millimeter wave (mmwave) communications. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, page 3630–3634, Mar 2016.
- [4] Olivier Bonaventure, Mohamed Boucadair, Sri Gundavelli, SungHoon Seo, and Benjamin Hesmans. 0-RTT TCP Convert Protocol. RFC 8803, July 2020.
- [5] David Borman. TCP Options and Maximum Segment Size (MSS). RFC 6691, July 2012.
- [6] Gwyn Chatranon, Miguel A. Labrador, and Sujata Banerjee. A survey of tcp-friendly router-based aqm schemes. *Computer Communications*, 27(15), September 2004.
- [7] Kristjon Ciko, Michael Welzl, and Peyman Teymoori. Pep-dna: A performance enhancing proxy for deploying network architectures. In *2021 IEEE 29th International Conference on Network Protocols (ICNP)*, pages 1–6, 2021.
- [8] J.-M. De Goyeneche and E.A.F. De Sousa. Loadable kernel modules. *IEEE Software*, 16(1):65–71, 1999.
- [9] Wesley Eddy. Transmission control protocol (tcp). Request for Comments RFC 9293, Internet Engineering Task Force, Aug 2022.

- [10] M. Garcia, J. Choque, L. Sanchez, and L. Munoz. An experimental study of snoop tcp performance over the ieee 802.11b wlan. In *The 5th International Symposium on Wireless Personal Multimedia Communications*, volume 3, page 1068–1072, Honolulu, HI, USA, 2002. IEEE.
- [11] S. Hingane. Using aqm to manage "temporary bufferbloat" on mmwave links. Run my experiment on GENI blog, 2020. Accessed: 2024-02-13.
- [12] Michio Honda, Yoshifumi Nishida, Costin Raiciu, Adam Greenhalgh, Mark Handley, and Hideyuki Tokuda. Is it still possible to extend tcp? In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, page 181–194, Berlin Germany, Nov 2011. ACM.
- [13] Toke Høiland-Jørgensen, Paul McKenney, dave.taht@gmail.com, Jim Gettys, and Eric Dumazet. *The Flow Queue CoDel Packet Scheduler and Active Queue Management Algorithm*. Number RFC 8290. Jan 2018.
- [14] Milosh Ivanovich, Philip W. Bickerdike, and Jonathan C. Li. On tcp performance enhancing proxies in a wireless environment. *IEEE Communications Magazine*, 46(9):76–83, 2008.
- [15] Brian W. Kernighan and Dennis M. Ritchie. *The C Programming Language*. Prentice Hall Professional Technical Reference, 2nd edition, 1988.
- [16] Euiyul Ko, Donghyeok An, Ikjun Yeom, and Hyunsoo Yoon. Congestion control for sudden bandwidth changes in tcp. *International Journal of Communication Systems*, 25(12):1550–1567, 2012.
- [17] James F. Kurose and Keith W. Ross. *Computer networking: a top-down approach*. Pearson, Boston, seventh edition edition, 2017.
- [18] Linus Torvalds and The Linux Kernel Developers. Linux kernel source code: include/linux/net.h.
- [19] Alberto Medina, Mark Allman, and Sally Floyd. Measuring interactions between transport protocols and middleboxes. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, page 336–341, Taormina Sicily, Italy, Oct 2004. ACM.
- [20] Michele Polese, Marco Mezzavilla, Menglei Zhang, Jing Zhu, Sundeep Rangan, Shivendra Panwar, and Michele Zorzi. milliproxy: A tcp proxy architecture for 5g mmwave cellular systems. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 951–957, 2017.

- [21] Michele Polese, Marco Mezzavilla, Menglei Zhang, Jing Zhu, Sundeep Rangan, Shivendra Panwar, and Michele Zorzi. milliproxy: a tcp proxy architecture for 5g mmwave cellular systems. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, page 951–957, October 2017. arXiv:1712.02700 [cs, math].
- [22] Matthew Rahtz. ultra_ping: A high precision ping tool. https://github.com/mrahtz/ultra_ping, 2023. Accessed: 2023-11-07.
- [23] Cristian García Ruiz, Antonio Pascual-Iserte, and Olga Muñoz. Analysis of blocking in mmwave cellular systems: Application to relay positioning. *IEEE Transactions on Communications*, 69(2):1329–1342, Feb 2021.
- [24] Seungwan Ryu, Christopher Rump, and Chunming Qiao. Advances in active queue management (aqm) based tcp congestion control. *Telecommunication Systems*, 25(3/4), March 2004.
- [25] Ashutosh Srivastava, Fraida Fund, and Shivendra S. Panwar. An experimental evaluation of low latency congestion control for mmwave links. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, page 352–357, Toronto, ON, Canada, July 2020. IEEE.
- [26] Viet-Hoang Tran and Olivier Bonaventure. Beyond socket options: making the linux tcp stack truly extensible. In *2019 IFIP Networking Conference (IFIP Networking)*, pages 1–9, 2019.
- [27] W. V. Wathsala, Buddhika Siddhisena, and Ajantha S. Athukorale. Next generation proxy servers. In *2008 10th International Conference on Advanced Communication Technology*, volume 3, pages 2183–2187, 2008.
- [28] Michael Welzl. Network congestion control: Managing internet traffic. *Network Congestion Control: Managing Internet Traffic*, pages 1–263, 05 2006.
- [29] Michael Welzl, Dimitri Papadimitriou, Bob Briscoe, Michael Scharf, and Michael Welzl. Open Research Issues in Internet Congestion Control. RFC 6077, February 2011.
- [30] Gina Yuan, Matthew Sotoudeh, David K. Zhang, Michael Welzl, David Mazières, and Keith Winstein. Sidekick: In-Network assistance for secure End-to-End transport protocols. In *21st USENIX Symposium on*

Networked Systems Design and Implementation (NSDI 24), pages 1813–1830, Santa Clara, CA, April 2024. USENIX Association.