

University of Groningen

P4-CoDel

Kundel, Ralf; Rizk, Amr; Blendin, Jeremias; Koldehofe, Boris; Hark, Rhaban; Steinmetz, Ralf

Published in:

Proceedings of the IEEE International Conference on Communications (ICC 2021): Next-Generation Networking and Internet Symposium

DOI:

[10.1109/ICC42927.2021.9500943](https://doi.org/10.1109/ICC42927.2021.9500943)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2021

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Kundel, R., Rizk, A., Blendin, J., Koldehofe, B., Hark, R., & Steinmetz, R. (2021). P4-CoDel: Experiences on Programmable Data Plane Hardware. In *Proceedings of the IEEE International Conference on Communications (ICC 2021): Next-Generation Networking and Internet Symposium* IEEE. <https://doi.org/10.1109/ICC42927.2021.9500943>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

P4-CoDel: Experiences on Programmable Data Plane Hardware

Ralf Kundel^{*}, Amr Rizk[¶], Jeremias Blending[‡], Boris Koldehofe[§], Rhaban Hark^{*}, Ralf Steinmetz^{*}

^{*} Multimedia Communications Lab, *Technical University of Darmstadt, Germany*

{ralf.kundel, rhaban.hark, ralf.steinmetz}@kom.tu-darmstadt.de

[¶] *University of Ulm, Germany*

amr.rizk@uni-ulm.de

[‡] Barefoot Networks, an Intel Company, *CA, US*

jeremias.blending@intel.com

[§] *University of Groningen, Netherlands*

b.koldehofe@rug.nl

Abstract—Fixed buffer sizing in computer networks, especially the Internet, is a compromise between latency and bandwidth. A decision in favor of high bandwidth, implying larger buffers, subordinates the latency as a consequence of constantly filled buffers. This phenomenon is called *Bufferbloat*. Active Queue Management (AQM) algorithms such as CoDel or PIE, designed for the use on software based hosts, offer a flow agnostic remedy to *Bufferbloat* by controlling the queue filling and hence the latency through subtle packet drops.

In previous work, we have shown that the data plane programming language P4 is powerful enough to implement the CoDel algorithm. While legacy software algorithms can be easily compiled onto almost any processing architecture, this is not generally true for AQM on programmable data plane hardware, i.e., programmable packet processors. In this work, we highlight corresponding challenges, demonstrate how to tackle them, and provide techniques enabling the implementation of such AQM algorithms on different high speed P4-programmable data plane hardware targets. In addition, we provide measurement results created on different P4-programmable data plane targets. The resulting latency measurements reveal the feasibility and the constraints to be considered to perform Active Queue Management within these devices. Finally, we release the source code and instructions to reproduce the results in this paper as open source to the research community.

Index Terms—CoDel, AQM, Bufferbloat, P4, ASIC, FPGA, NPU

I. INTRODUCTION

Bufferbloat describes a phenomenon of high latencies observed in networks configured with large static buffer sizes [1]. For a single TCP traffic flow, it is known that the buffer size that maximizes the TCP throughput is directly proportional to the round trip time (RTT). This is also true for the case of multiplexing many homogeneous flows having the same RTT except of a correction prefactor [2]. As traffic flows usually have widely different RTTs and throughput maximization is not the main goal for many contemporary networking applications, but rather latency minimization, an idea of controlling flow delays through subtle packet drops is experiencing a renaissance [3]. The algorithmic version of this idea is denoted Active Queue Management (AQM) and is based on the sensitive reaction of the transport protocol congestion control, typically a TCP

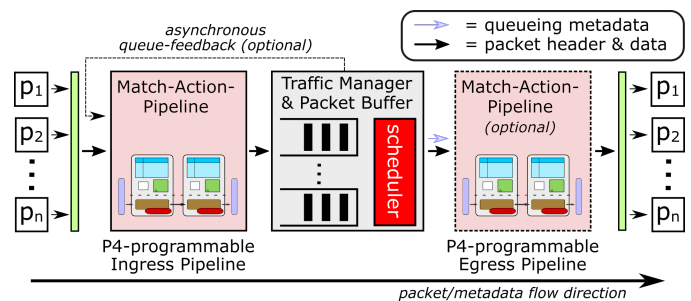


Fig. 1: Functional building blocks of P4-programmable ASICs. The light-red parts are P4 programmable. Packet queueing is not part of P4 and its configuration is vendor dependent. A programmable Match-Action Pipeline after the packet buffer is not given in all currently existing architectures.

variant, to packet drops. By dropping packets earlier than at a full buffer, the sender congestion control algorithm receives an *early signal* to reduce the sending data rate. In turn, this leads to the buffer filling and, hence, the *queueing delay* to remain relatively small. In recent years, two stateful and self-tuning AQM algorithms, i.e., CoDel (RFC 8289) [4] and PIE (RFC 8033), have been presented and widely adopted. In addition to those two approaches, there exist many variants of AQM algorithms with different goals and assumptions [5].

In a previous work, we have shown that such AQM algorithms can be expressed with programming languages aimed at controlling network switch data plane behavior [6]. However, it remained open how feasible it is to realize such AQM algorithms on packet processors, as these algorithms were not conceptualized to run on data plane network hardware but rather on software based consumer edge devices. Indeed, for many networking applications packet processing with high throughput is required and can only be ensured by algorithms realized directly within the data plane. Further, it remained open how different programmable networking hardware affect the algorithms performance and which additional challenges arise in different architectures.

To understand the problem of running AQM algorithms on programmable data plane hardware, a deeper look into the pipeline of modern packet processors is required. The internal

pipeline of packet processors, including network switches, is typically similar to the architecture depicted in Figure 1. Packets enter on one of the n ports on the left side and are multiplexed into a single match-action-pipeline. Within this pipeline operations, e.g. lookups, on packet header fields and metadata can be performed. After that, the traffic manager is responsible for packet queuing and scheduling. Optionally, depending on the architecture, a second match-action-pipeline can be applied on the packets before demultiplexing and sending the packets out on the specified egress port. Note that packets can flow only from left to right and *algorithmic loops within the pipeline are not possible*. By that, a high and deterministic processing performance can be guaranteed which is crucial for network functions running at line rate [7], including AQM algorithms running within the network.

In case of programmable packet processors built on top of the Protocol Independent Switch Architecture (PISA) [8], these ingress and egress match-action-pipelines are programmable within limitations. The programming language P4 [9] represents the currently most widely accepted approach of programming the depicted ingress and egress match-action-pipelines. However, the configuration of the packet buffer in the middle of the switch architecture (see Figure 1) as well as the specific algorithms used by this engine for queuing, scheduling and buffer management are out of the scope and purpose of this language. Nevertheless, useful metadata produced inside the Traffic manager, such as the packet queuing delay, can be passed alongside with other packet metadata into the egress pipeline. Alternatively, depending on the actual architecture, the current buffer utilization can be passed asynchronously to the ingress pipeline.

The work at hand focuses on the feasibility of realizing AQM algorithms on programmable hardware devices, which creates a variety of challenges compared to the implementations on classical commodity processor. Here, we analyze the required alterations of an established stateful AQM algorithm (CoDel) that make it possible to implement it. Note that our key findings are not dependent on the specific AQM algorithm at hand. We evaluate P4-Codel for P4-NetFPGAs [10], Neronome SmartNICs and Intel Tofino switches.

The main contributions of this paper are:

- The analysis of CoDel’s AQM implementation control flow, particularly considering information dependencies,
- how to transform an algorithm designed for CPUs for packet processing pipelines and a detailed implementation description of CoDel for existing P4 hardware,
- evaluation results of characteristic properties for these different hardware implementations,
- an open source implementation of the presented algorithm for two different hardware architectures including reproduction instructions.

II. THE CODEL AQM ALGORITHM

In the following, we use the CoDel AQM-algorithm presentation given in Listing 1 to illustrate the stateful data dependencies within the algorithm. The effectiveness of CoDel-

algorithm, presented in 2018 by Nichols *et al* [4], will not be discussed further. Figure 2 shows the control flow graph of the algorithm that includes four stateful variables: 1) *dropping*, 2) *count*, 3) *last_count* and 4) *drp_next*. The value of a stateful variable persist the processing time of a single packet.

The algorithm can be regarded as state machine consisting of two states: 1) *dropping* and 2) *not dropping*. As soon as the observed queuing delay exceeds the *TARGET* of 5 ms, the state of the state machine is changed to *dropping* which, however, does not imply an immediate packet drop (*if_2*). After waiting for the time *INTERVAL* until time *drp_next* within the *dropping* state, the first following packet is dropped and the *counter* of dropped packets is increased (*if_4*). From this point on, the interval between dropping packets is continuously decreased as $\frac{INTERVAL}{\sqrt{count}}$ until the queuing delay falls below the *TARGET* delay. Then, the state changes back to *not dropping* (*if_1*). In case of a recently left *dropping* state, the new dropping rate is initialized higher than 1 (*if_3*).

The CoDel algorithm is conceptualized for sequential packet processing, i.e., no parallel processing of multiple packets. Thus, the processing of *packet_n* is expected to be completed before the processing of *packet_{n+1}* starts. Otherwise, e.g., the read operation on the stateful variable *drp_next* for *packet_{n+1}* is performed before the write operation on this variable by *packet_n* is completed and by that unexpected behavior occurs.

Note that from an algorithmic perspective a partial overlapping processing of multiple packets is possible. If for each stateful variable the write operation is performed before the read operation of the subsequent packet, the algorithm is executed correctly. In the concrete case of CoDel, operations on

```

1 #define TARGET 5 //ms
2 #define INTERVAL 100 //ms
3 Queue q; State s;
4
5 upon receive packet p:
6 //no target delay violation? (if_1)
7 if(p.queue_delay < TARGET || q.byte < IFACE_MTU):
8     s.dropping = false
9     continue
10 //first packet which violates delay target? (if_2)
11 if(s.dropping == false):
12     s.dropping = true
13     tmp = s.count
14     //drop frequency steady state? (if_3)
15     if((s.count - s.last_count > 1) &&
16        (now - s.drp_next_packet < 16*INTERVAL)):
17         //optimization for faster packet dropping
18         s.count = s.count - s.last_count
19     else:
20         s.count = 1
21     s.drp_next_packet = now + INTERVAL/sqrt(s.count)
22     s.last_count = tmp
23     continue
24 //drop state for at least x time units? (if_4)
25 if(s.dropping && s.drp_next_packet <= now):
26     p.drop()
27     s.count++
28     s.drp_next_packet = now + INTERVAL/sqrt(s.count)

```

Listing 1: CoDel Pseudocode, based on RFC 8289 [4].

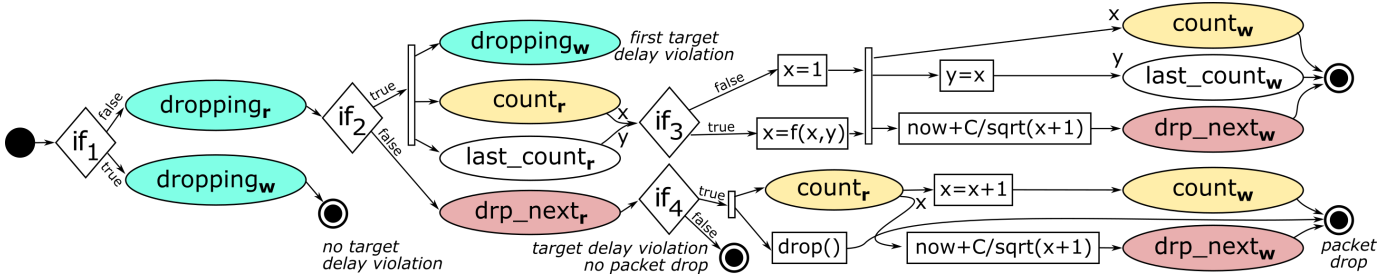


Fig. 2: General stateful data centric program flow of the CoDel-algorithm. Read access on stateful variables is indicated by r , write operations by w .

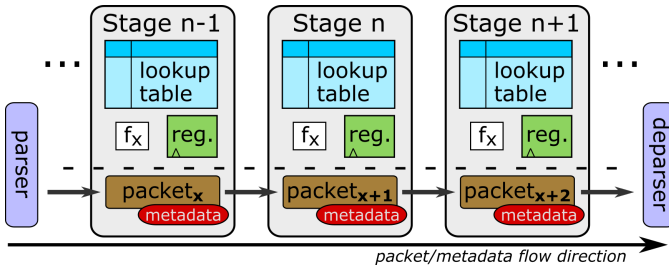


Fig. 3: Generic acyclic P4-programmable Match-Action Pipeline including packet (de) parsing. Periodically, after each pipeline clock cycle, all packets are forwarded by one stage. Backwards information flow is not possible. Each stateful information is attached to a register in one dedicated stage.

the stateful *dropping* variable can be considered independently to the other three state variables and by that executed in parallel for consecutive packets. As noticeable from Figure 2, the other three stateful variables cannot be isolated as the operations depend on each other. Thus, all operations on these three variables must be executed as an atomic block.

The structure of programmable packet pipelines, depicted in Figure 3, does not allow cyclic data dependencies over multiple stages. This means, a stateful variable of the algorithm located in a register in stage $n - 1$ can only be read there and used for further computations in stage $n - 1$, n and $n + 1$. In this case, however, a read-compute-write cycle is only possible within stage $n - 1$ as the result cannot be fed back from a later stage to stage $n - 1$ which is then already processing the subsequent packet. As a consequence, each atomic stateful operation must be executed within one single pipeline stage including reading and writing all required variables. We observe that other AQM algorithms such as PIE can be analyzed in a similar way in order to obtain the stateful data dependencies.

III. HARDWARE IMPLEMENTATION

In the following, we describe three efforts to migrate and run the CoDel algorithm on significantly different P4-programmable hardware platforms. Recall that P4 is designed to describe packet header processing and not for queueing.

A. PISA-architecture switches

The PISA platform allows the ingress and egress pipeline, consisting each of N stages, to be programmed with P4. As the queueing delay information is only available within the egress pipeline we decided to implement CoDel there. In addition,

the traffic manager needs to be configured accordingly to a constant rate, e.g., 100 Mbit/s, in order to build up a queue.

As stated before, the algorithm must be loop- and cycle-free over different pipeline stages and information can only flow from the left to the right as shown in Figure 3. In order to achieve this, we adapt the algorithm as shown in Listing 2. This adaptation targets the mapping of each atomic operation onto a single match-action pipeline stage. Note that in case of Intel Tofino, following the PISA-architecture, allows the P4 program to execute a small number of different operations, such as f_x , on the contents of the register within a single stage using a special ALU, called stateful ALU (see Figure 3).

First, we determine whether the *TARGET* delay is violated or not in function f_1 . This operation is stateless and can be performed for each packet independently. Second, we update the *dropping* state and deviate if the current delay violation is not preceded by a delay violation, which means the state is changing from *non_dropping* to *dropping*. Third, the decision whether a packet should be dropped, the computation of the next drop time (*drp_next*) and incrementing the drop *counter* are performed within another stateful ALU. The output of this ALU, either 0 or 1, indicates if the current packet should be dropped if the *TARGET* delay is violated as well (*codel_drop*). The square root function can be approximated very well by the math unit within the stateful ALU. Last, the final dropping decision is performed by checking if the *TARGET* delay is violated (output of f_1) and the second stateful ALU marks this packet to be dropped.

The mapping of these functional blocks on the PISA pipeline is depicted in Figure 4. In total, the algorithm requires four pipeline stages: 1) performing the f_1 computation, 2,3) stateful ALU 1 and 2, and 4) the final drop decision.

Compared to the algorithm in Listing 1, the optimization for more frequent packet dropping in the beginning which is marked as *if_3* could not be implemented on this PISA hardware. The reason for that is, that the number of registers and arithmetic operations that can be performed of them by a single stateful ALU is limited in order to provide real time guarantees and high performance up to the line rate of 100Gbit/s per port for many ports in parallel. Furthermore, for newer hardware generations we expect even more flexible stateful ALUs and, as shown later in evaluation section, the impact of this limitation on the performance is very limited.

Fq_CoDel: An evolution of CoDel is given by fq_CoDel

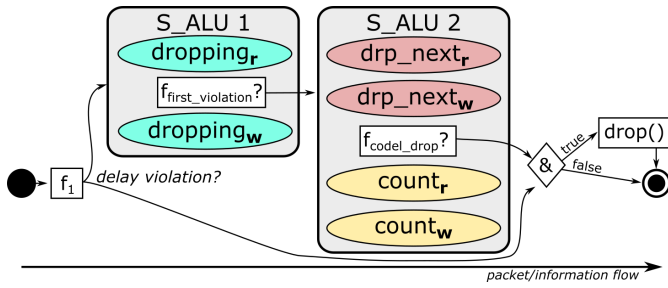


Fig. 4: Stateful data centric program flow of the CoDel-algorithm rewritten without cyclic data dependencies optimized for the PISA architecture. Each gray box represents a stateful ALU which guarantees atomic execution within one single pipeline stage.

which can be realized on PISA platforms as well. This algorithm distributes incoming packets in the ingress pipeline over multiple queues for one single egress port. These queues are served by a (weighted) round robin scheduler and for each queue a CoDel state is required in the egress pipeline. As the stateful ALU provide an addressable register set instead of a single register, this can be realized as well.

B. NPU-based packet processors

The internal structure of so called Network Processing Units (NPU) is similar to a many-core processor or GPU. Incoming packets are divided over a set of processors, that execute in parallel the compiled program code of the P4-programmable ingress pipeline. After that, the packets are enqueued in per port egress queues with a certain rate limit, e.g., 100 Mbit/s. After these queues, a processing by a P4 programmable block is not possible any more. However, as shown in Figure 1, an asynchronous feedback from the packet queues to the P4-programmable ingress pipeline is possible. Thus, the AQM has to be implemented within the ingress pipeline and the input to the algorithm is the current delay of the queue and not the queueing delay of the current packet which are only subtly different. We noticed that access to this information on the queue level is only possible within a microC-sandbox, executed as external P4 function, and not directly in P4. Due to further complexities while handing over data from and to this sandbox, we decided to realize the CoDel AQM as a fixed P4-external function within a sandbox.

```

1 #define TARGET 5 //ms
2 #define INTERVAL 100 //ms
3 Packet p; Queue q;
4 //check for target delay violation? (f_1)
5 if(p.queue_delay < TARGET || q.byte < IFACE_MTU):
6     delay_violation = false
7 else:
8     delay_violation = true
9     first_violation = S_ALU1.exec(delay_violation)
10    codel_drop = S_ALU2.exec(first_violation)
11    if(delay_violation && codel_drop):
12        p.drop()

```

Listing 2: Rewritten CoDel algorithm for fitting into loop- and -free P4-programmable PISA pipelines. This code does not contain all implementation details and just give, together with Figure 4 an overview.

The approximation of the square root function was realized by an exact lookup table for the values $count = 1..16$ and by an approximation for all other values. This approximation, already presented in [6], uses a longest prefix matcher in order to count the leading zeros of the stateful $count$ variable.

On this hardware architecture we observed two main issues: First, as multiple parallel processing units are running within the SmartNIC, race conditions between multiple accesses on stateful variables can occur. Due to the internal processor architecture, there is no trivial solution to this challenge. Second, to the best of our knowledge, the queue depth can only be measured by the number of enqueued packets which is, in case of heterogeneous packet sizes, a major constraint.

C. P4-NetFPGA

Lastly, we investigated the P4-NetFPGA architecture [10]. This architecture provides only tiny queues per egress ports. In addition, these queues are not rate limited and always send packets out with the link rate, i.e., in this case 10Gbit/s. However, we analyzed possibilities of getting CoDel to work within the P4 pipeline. As the stateful operations turned out to be challenging in this context, we decided to provide the AQM algorithm as a P4 external function to the pipeline, similar to the NPU-based SmartNIC approach given above in Sect. III-B. Such an external AQM algorithm, however, has to be designed in a low level programming language as Verilog or VHDL. Finally, our running CoDel-prototype for P4-NetFPGA is avoiding all components of the P4-pipeline and all AQM-crucial components are realized as standalone modules. Consequently, our implementation would run even without a P4 framework and therefore we do not show results for P4-NetFPGA in the following.

IV. EVALUATION

The following evaluation experiments are performed using a testbed built upon the P4STA load aggregation and measurement framework [11] as depicted in Figure 6. The Device Under Test (DUT), encircled by the P4STA-Stamper for high accuracy time measurements and loss detection, represents the CoDel implementation on a P4-pipeline realized by (i) the Intel Tofino ASIC, (ii) Netronome NPU-based SmartNICs and the (iii) Linux kernel as reference implementation. All scripts for reproducing the results are available as open source software together with the CoDel source code ¹.

Considering Figure 6 each packet sent by any of the three TCP senders on the left side is timestamped and counted before entering the CoDel implementation of the DUT. The same is performed for each packet leaving on the right side to the receivers and by that very accurate queueing delay measurements are performed. TCP acknowledgments traverse the setup back smoothly without any packet loss.

In a first run, a single TCP sender and receiver, using iPerf3, transfer as much data as possible over the bottleneck link. The results for the three investigated CoDel AQM DUTs are shown

¹<https://github.com/ralfkundel/p4-codel>

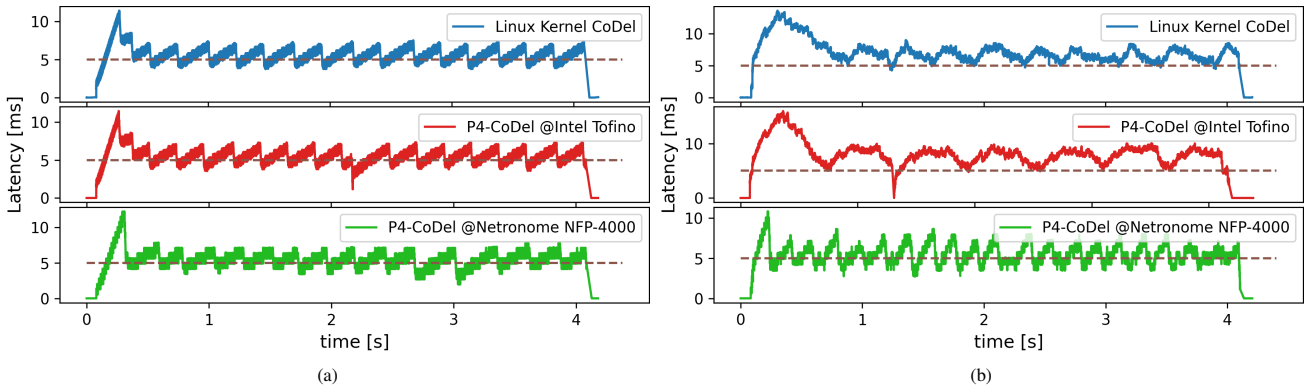


Fig. 5: Measured latency for CoDel realizations on P4-programmable packet switching ASICs, NPU-based P4-SmartNICs, vs the Linux kernel reference implementation for (a) one and (b) three parallel TCP flows (cf. Fig. 6 for measurement setup).

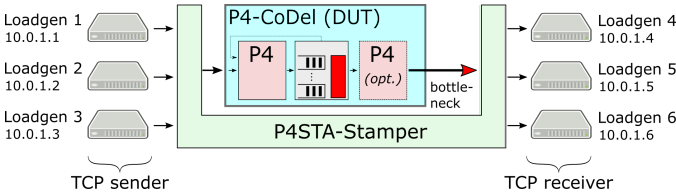


Fig. 6: Data plane testbed setup based on the P4STA load aggregation framework [11]. Each server can create multiple parallel TCP flows. Each server can be assigned an additional link latency for heterogeneous RTTs.

in Figure 5a. In all three scenarios an initial burst of packets is filling the buffer to a latency slightly above $10ms$ until CoDel reacts and reduces the latency by subtle packet drops which results in a reduced TCP sending rate. By that, the queuing delay falls periodically below the configured *TARGET* delay of $5ms$. On the rising edge one may notice a oscillation of the latency. This behavior can be explained by a microburstiness of the TCP sender [12]. In Figure 5b we consider three senders in parallel where this patterns becomes noisy. For the NPU-based Netronome SmartNICs this noise is slightly higher; this could be caused by the internal many-core processor architecture, however, we could neither confirm nor refute this.

In addition to considering only one single TCP flow, we will focus next on the case of multiplexing multiple flows. Table I shows the number of dropped packets by the CoDel implementation and the average latency for 1,2 and 3 parallel TCP flows. First, we notice that the number of dropped packets is strongly increasing with the number of parallel TCP flows. This is caused by the fact that CoDel drops only one single packet and by that the corresponding TCP flow is reducing its sending rate but the other flows benefit from the released free link capacity. As a consequence, an AQM in general has to drop more packets in order to control all flows and by that the latency. Second, the average latency increase on the P4-ASIC when going from 1 to 3 TCP flows is significant. This can be explained by the missing optimization (*if_3*) for this P4 target.

In Figure 5b we show the latency for the three TCP flows where we observe a similar behavior for the P4-ASIC and the Linux kernel. The results for the NPU-based SmartNICs suggest a TCP flow synchronization which is usually avoided by the CoDel algorithm. Lastly note that the drop rate of the

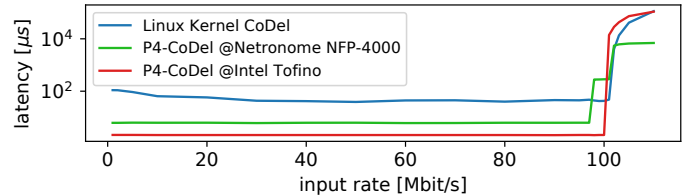


Fig. 7: Average latency for isochronous shaped UDP traffic (no congestion control) with a varying rate and a CoDel rate limit of $100Mbit/s$.

Linux kernel is slightly lower. The reason behind this could be the difference between the kernel implementation and the RFC implementation we build upon.

Figure 7 depicts the observed average latency (mainly due to CoDel processing on different hardware) for a rate scan with shaped constant rate UDP streams. Note that the CoDel shaping rate is $100Mbit/s$, i.e. the latency is expected to start increasing around that point due to queue filling within the DUT. The Linux implementations shows an increased average latency for small data rates due to the Linux kernel interrupt behavior. The latency of Netronome SmartNICs starts increasing at $97Mbit/s$, however, a persistent queue building starts as expected at rate higher than the CoDel shaping rate.

V. RELATED WORK

Programmable queueing and scheduling algorithms have been discussed, e.g., in Sharma *et al* [13] who proposed in 2018 a queueing mechanism, called “Approximate Fair Queueing”, prioritizing packets in the programmable data plane in order to achieve shorter flow completion times. In a follow up work they proposed the idea of programmable queues based on the construct of calendar queues [14] which provide high flexibility in programming schedulers. However, this approach relies on queues which can be re-parameterized from within the data plane which is not supported by existing switches.

#flows	Linux		P4-NPU		P4-ASIC	
	loss	latency	loss	latency	loss	latency
1	0.05%	5.7 ms	0.17 %	5.51 ms	0.16 %	5.59 ms
2	0.17%	6.5 ms	0.24 %	6.25 ms	0.36 %	6.86 ms
3	0.33%	7.2 ms	0.45 %	6.42 ms	0.44 %	8.17 ms

TABLE I: Observed packet drops and latency for the different investigated targets. Each run is $4s$ with a rate of $100Mbit/s$ and $33 * 10^3$ packets.

Another approach of programming queues within switches is the approximation of Push-In-First-Out (PIFO) queues by conventional FIFO queues. Alcoz *et al* [15] presented a P4-based implementation, called “SP-PIFO”, of this approach running in existing data planes with a behavior that is very close to ideal PIFO queues. The authors of “ConQuest” [16] tackled the impact of short-lived packet surges by monitoring the network traffic from within the data plane and apply some queue management. Specifically, similar to our work, they present a prototype that is based on a P4-programmable Intel Tofino switch which is able to identify these surges and avoid congestion by early marking or dropping certain packets.

In a previous work we have shown the feasibility of CoDel-AQM algorithm in the programming language P4 [6]. Based on that, Papagianni *et al* have introduced “PI2 for P4” [17] that provides a P4 implementation of the AQM algorithm PI2 [18] for programmable data planes.

VI. CONCLUSION

Active Queue Management (AQM) algorithms enable high throughput and low latency simultaneously and provide a practical solution to the *Bufferbloat* phenomenon. In this work, we have shown how recent self-tuning AQM algorithms can be implemented on P4-programmable hardware without built-in queue management capabilities achieving much better performance than in software systems. Further, we discussed different challenges that arise depending on the chosen hardware platform to run the AQM algorithm. While the programming language P4 itself is quite powerful, writing compilable and correct behaving AQM algorithms leads to challenges due to hardware constraints. With this in mind, we note that correct hardware implementations, are far superior in terms of speed and deterministic behavior to software implementations.

We investigated multiple P4 target platforms with quite diverse internal architectures: The PISA architecture, concrete the Intel Tofino ASIC, is very restrictive on the one hand, but turned out to be the most powerful in terms of performance one on the other hand. The investigated NPU-based SmartNICs can be programmed in a very flexible way but facilitate causing race conditions within the developed AQM-program. Further, the P4-NetFPGA architecture does not support large packet buffers and hinders a successful implementation from an algorithmic point of view. Only by strong modifications outside the P4-pipeline this architecture is able to perform AQM, subverting the idea of this work.

Finally, our results show that programmable data plane hardware is not only suitable for header processing but also for flexible queue management. We noticed that advanced queue management algorithms require stateful processing capabilities in the underlying hardware. To enable a seamless integration of such algorithms, we hope with this paper to start a discussion on AQM interface definitions for P4 programmable pipelines.

ACKNOWLEDGMENT

This work has been supported by Deutsche Telekom through the Dynamic Networks 8 project, and in parts by the German

Research Foundation (DFG) as part of the projects B4 and C2 within the Collaborative Research Center (CRC) 1053 MAKI as well as the DFG project SPINE. Furthermore, we thank our colleagues and reviewers for their valuable input and feedback.

REFERENCES

- [1] J. Gettys and K. Nichols, “Bufferbloat: dark buffers in the internet,” *Communications of the ACM*, vol. 55, no. 1, pp. 57–65, 2012.
- [2] G. Appenzeller, I. Keslassy, and N. McKeown, “Sizing router buffers,” in *Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*. ACM, 2004, p. 281–292.
- [3] S. Floyd and V. Jacobson, “Random early detection gateways for congestion avoidance,” *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, 1993.
- [4] K. Nichols, V. Jacobson, A. McGregor, and A. Iyengar, “Controlled Delay Active Queue Management,” IETF, RFC 8289, 2018.
- [5] A. Sivaraman, K. Winstein, S. Subramanian, and H. Balakrishnan, “No silver bullet: extending sdn to the data plane,” in *Proceedings of the Twelfth ACM Workshop on Hot Topics in networks*. ACM, 2013, p. 19.
- [6] R. Kundel, J. Blendin, T. Viernickel, B. Koldehofe, and R. Steinmetz, “P4-codel: Active queue management in programmable data planes,” in *Proceedings of the Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*. IEEE, 2018, pp. 1–4.
- [7] R. Kundel, L. Nobach, J. Blendin, W. Maas, A. Zimmer, H.-J. Kolbe, G. Schyguda, V. Gurevich, R. Hark, B. Koldehofe, and R. Steinmetz, “OpenBNG: Central office network functions on programmable data plane hardware,” *International Journal of Network Management*, 2021.
- [8] P. Bosshart, G. Gibb, H.-S. Kim, G. Varghese, N. McKeown, M. Izzard, F. Mujica, and M. Horowitz, “Forwarding metamorphosis: Fast programmable match-action processing in hardware for sdn,” in *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, 2013.
- [9] P. Bosshart, D. Daly, G. Gibb, M. Izzard, N. McKeown, J. Rexford, C. Schlesinger, D. Talayco, A. Vahdat, G. Varghese *et al.*, “P4: Programming protocol-independent packet processors,” *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 3, pp. 87–95, 2014.
- [10] S. Ibanez, G. Brebner, N. McKeown, and N. Zilberman, “The p4 → netfpga workflow for line-rate packet processing,” *Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, 2019, pp. 1–9.
- [11] R. Kundel, F. Siegmund, J. Blendin, A. Rizk, and B. Koldehofe, “P4STA: High performance packet timestamping with programmable packet processors,” in *Network Operations and Management Symposium (NOMS)*. IEEE/IFIP, 2020, pp. 1–9.
- [12] R. Kundel, A. Rizk, and B. Koldehofe, “Microbursts in software and hardware-based traffic load generation,” in *Network Operations and Management Symposium (NOMS)*. IEEE/IFIP, 2020, pp. 1–2.
- [13] N. K. Sharma, M. Liu, K. Atreya, and A. Krishnamurthy, “Approximating fair queueing on reconfigurable switches,” in *Symposium on Networked Systems Design and Implementation*. USENIX, 2018.
- [14] N. K. Sharma, C. Zhao, M. Liu, P. G. Kannan, C. Kim, A. Krishnamurthy, and A. Sivaraman, “Programmable calendar queues for high-speed packet scheduling,” in *17th Symposium on Networked Systems Design and Implementation (NSDI)*. Santa Clara, CA: USENIX Association, Feb. 2020, pp. 685–699.
- [15] A. G. Alcoz, A. Dietmüller, and L. Vanbever, “Sp-pifo: Approximating push-in first-out behaviors using strict-priority queues,” in *17th Symposium on Networked Systems Design and Implementation*. USENIX, 2020, pp. 59–76.
- [16] X. Chen, S. L. Feibish, Y. Koral, J. Rexford, O. Rottenstreich, S. A. Monetti, and T.-Y. Wang, “Fine-grained queue measurement in the data plane,” in *Proceedings of the 12th International Conference on Emerging Networking EXperiments and Technologies*. ACM, 2019, pp. 15 – 29.
- [17] C. Papagianni and K. De Schepper, “Pi2 for p4: An active queue management scheme for programmable data planes,” in *Proceedings of the 15th International Conference on Emerging Networking EXperiments and Technologies*. ACM, 2019, pp. 84–86.
- [18] K. De Schepper, O. Bondarenko, I.-J. Tsang, and B. Briscoe, “Pi2: A linearized aqm for both classic and scalable tcp,” in *Proceedings of the 12th International Conference on Emerging Networking EXperiments and Technologies*. ACM, 2016, pp. 105–119.