

Piece of CAKE: A Comprehensive Queue Management Solution for Home Gateways

Toke Høiland-Jørgensen
Dept. of Computer Science
Karlstad University, Sweden
toke.hoiland-jorgensen@kau.se

Dave Täht
Teklibre
Los Gatos, California
dave.taht@gmail.com

Jonathan Morton
Somero, Finland
chromatix99@gmail.com

Abstract—The last several years has seen a renewed interest in smart queue management to curb excessive network queueing delay, as people have realised the prevalence of bufferbloat in real networks.

However, for an effective deployment at today’s last mile connections, an improved queueing algorithm is not enough in itself, as often the bottleneck queue is situated in legacy systems that cannot be upgraded. In addition, features such as per-user fairness and the ability to de-prioritise background traffic are often desirable in a home gateway.

In this paper we present *Common Applications Kept Enhanced (CAKE)*, a *comprehensive network queue management system* designed specifically for home Internet gateways. CAKE packs several compelling features into an integrated solution, thus easing deployment. These features include: bandwidth shaping with overhead compensation for various link layers; reasonable DiffServ handling; improved flow hashing with both per-flow and per-host queueing fairness; and filtering of TCP ACKs.

Our evaluation shows that these features offer compelling advantages, and that CAKE has the potential to significantly improve performance of last-mile internet connections.

I. INTRODUCTION

Eliminating bufferbloat has been recognised as an important component in ensuring acceptable performance of internet connections, especially as applications and users demand ever lower latencies. The last several years have established that Active Queue Management and Fairness Queueing are effective solutions to the bufferbloat problem, and several algorithms have been proposed and evaluated (e.g., [1]–[3]).

However, while modern queueing algorithms can effectively control bufferbloat, effective deployment presents significant challenges. The most immediate challenge is that the home gateway device is often not directly in control of the bottleneck link, because queueing persists in drivers or firmware of devices that cannot be upgraded [1]. In addition, other desirable features in a home networking context (such as per-user fairness, or the ability to explicitly de-prioritise background applications) can be challenging to integrate with existing queueing solutions. To improve upon this situation, we have developed *Common Applications Kept Enhanced (CAKE)*, which is a *comprehensive network queue management system* designed specifically for the home router use case.

As outlined below, each of the issues that CAKE is designed to handle has been addressed separately before. As such, the

compelling benefit of CAKE is that it takes state of the art solutions and integrates them to provide:

- a high-precision rate-based bandwidth shaper that includes overhead and link layer compensation features for various link types.
- a state of the art fairness queueing scheme that simultaneously provides both host and flow isolation.
- a Differentiated Services (DiffServ) prioritisation scheme with rate limiting of high-priority flows and work-conserving bandwidth borrowing behaviour.
- TCP ACK filtering that increases achievable throughput on highly asymmetrical links.

CAKE is implemented as a *queueing discipline* (qdisc) for the Linux kernel. It has been deployed as part of the OpenWrt router firmware for the last several years and is in the process of being submitted for inclusion in the mainline Linux kernel.¹

The rest of this paper describes the design and implementation of CAKE and is organised as follows: Section II outlines the desirable features of a comprehensive queue management system for a home router, and recounts related work in this space. Section III describes the design and implementation of CAKE in more detail, and Section IV evaluates the performance of the various features. Finally, Section V concludes.

II. BACKGROUND AND RELATED WORK

As mentioned initially, CAKE is designed to run on a home network gateway. We have gathered significant experience with implementing such a system in form of the Smart Queue Management (SQM) system shipped in the OpenWrt router firmware project, which has guided the design of CAKE.

In this section we provide an overview of the problems CAKE is designed to address. We are not aware of any previous work addressing the home gateway queue management challenges as a whole. However, several of the issues that CAKE addresses have been subject of previous work, and so the following subsections serve as both an introduction to the design space and an overview of related work.

The four problems we seek to address are bandwidth shaping, queue management and fairness, DiffServ handling and TCP ACK filtering. These are each treated in turn in the following sections.

¹We include links to the source code, along with the full evaluation dataset, in an online appendix [4].

A. Bandwidth Shaping

A queue management algorithm is only effective if it is in control of the bottleneck queue. Thus, queueing in lower layers needs to be eliminated, which has been achieved in Linux for Ethernet [5] and WiFi [6]. However, eliminating queueing at the link layer is not always possible, either because the driver source code is unavailable, or because the link-layer is implemented in inaccessible hardware or firmware (either on the same device or a separate device, such as a DSL modem).

As an alternative, queueing in the lower layers can be avoided by deploying a bandwidth shaper as part of the queue management system. By limiting the traffic traversing the bottleneck link to a bandwidth that is slightly less than the physical capacity of the link itself, queueing at the physical bottleneck can be eliminated and bufferbloat avoided. Such bandwidth shaping can be performed by a token bucket-based shaper (as is well-known from ATM networks, e.g., [7]), or by a rate-based shaper (which is known from video streaming applications, e.g., [8]).

The use of a shaper to move the link bottleneck wastes the bandwidth that is the difference between the actual physical link capacity, and the set-point of the shaper. To limit this waste, the shaper needs to be set as close to the actual link bandwidth as possible, while avoiding sending bursts of packets at a rate that is higher than the actual capacity. To achieve this, accurate timing information on a per-packet basis is needed. In addition, the shaper must account for link-layer framing and overhead. For instance, DSL links using ATM framing split up data packets into an integer number of fixed-size cells, which means that the framing overhead is a step function of packet size, rather than a fixed value.

B. Queue Management

Having control of the bottleneck queue makes it possible to implement effective queue management that can all but eliminate bufferbloat. Such a queue management scheme usually takes the form of an Active Queue Management (AQM) algorithm, combined with a form of fairness queueing (FQ). Several such schemes exist, and extensive evaluation is available in the literature (e.g., [1]–[3], [9]–[11]).

Among the state of the art algorithms in modern queue management, is the FQ-CoDel algorithm [12]. FQ-CoDel implements a hybrid AQM/fairness queueing scheme which isolates flows using a hashing scheme and schedules them using a Deficit Round-Robin (DRR) [13] scheduler. In addition, FQ-CoDel contains an optimisation that provides implicit service differentiation for sparse (low-bandwidth) flows, similar to [14], [15]. Evaluations of FQ-CoDel have shown that it achieves low queueing latency and high utilisation under a variety of scenarios [1], [3].

However, while the FQ-CoDel scheduler provides flow isolation and fairness, the transport layer flow is not always the right level of fairness in the home gateway use case. Often, additional isolation between *hosts* on the network is desirable; and indeed this per-host isolation was the most requested feature of the SQM system. Host isolation is straight-forward

to implement in place of flow fairness in any fairness queueing based scheme (by simply changing the function that maps packets into different queues), but we are not aware of any practical schemes prior to CAKE that implement *both* host and flow fairness.

C. DiffServ Handling

Even though flow-based fairness queueing offers a large degree of separation between traffic flows, it can still be desirable to explicitly treat some traffic as higher priority, and to have the ability to mark other traffic as low priority. Since a home network generally does not feature any admission control, any prioritisation scheme needs to be robust against attempts at abuse (so, e.g., a strict priority queue does not work well). In addition, enabling prioritisation should not affect the total available bandwidth in the absence of marked traffic, as that is likely to cause users to turn the feature off.

Prioritisation of different traffic classes can be performed by reacting to DiffServ markings [16]. This is commonly used in WiFi networks, where DiffServ code points map traffic into four priority levels [17]. For the home gateway use case, various schemes have been proposed in the literature (e.g., [18]), but as far as we are aware, none have seen significant deployment.

D. TCP ACK Filtering

TCP ACK filtering is an optimisation that has seen some popularity in highly asymmetrical networks [19], and especially in cable modem deployments [20]. The technique involves filtering (or *thinning*) TCP acknowledgement (ACK) packets by inspecting queues and dropping ACKs if a TCP flow has several consecutive ACKs queued. This can improve performance on highly asymmetrical links, where the reverse path does not have sufficient capacity to transport the ACKs produced by the forward path TCP flow. However, ACK filtering can also have detrimental effects on performance, for instance due to cross layer interactions [21].

III. THE DESIGN OF CAKE

The design of CAKE builds upon the basic fairness scheduler design of FQ-CoDel, but adds features to tackle the areas outlined in the previous section. The following sections outline how CAKE implements each of these features.

A. Bandwidth Shaping

CAKE implements a rate-based shaper, which works by scheduling packet transmission at precise intervals using a virtual transmission clock. The clock is initialised by the first packet to arrive at an empty queue, and thereafter is incremented by the calculated serialisation delay of each transmitted packet. Packets are delayed until the system time has caught up with the virtual clock. If the clock schedule is reached while the queue is empty, the clock is reset and the link goes idle.

This shaper handles bandwidth ranging over several orders of magnitude, from several Kbps to several Gbps. In addition,

the rate-based shaper does not require a burst parameter, which simplifies configuration as compared to a token-bucket shaper. It also eliminates the initial burst observed from token-bucket shapers after an idle period. This is important for controlling the bottleneck queue, as this initial burst would result in queueing at the real bottleneck link.

1) *Overhead and Framing Compensation:* As mentioned in Section II-A above, the shaper accounts for the actual size of a packet on the wire, including any encapsulation and overhead, which allows the rate to be set closer to the actual bottleneck bandwidth, thus eliminating waste. We believe it is safe to set a rate within 0.1% of the actual link rate when the overhead compensation is configured correctly, with a margin mainly required to accommodate slight variations in the actual bottleneck link bandwidth, caused by, e.g., clock drift in the hardware.

CAKE implements an overhead compensation algorithm which begins by determining the size of the network-layer packet, stripped of any MAC layer encapsulation. Having determined the network-layer packet size, the configured overhead can be added to yield the correct on-the-wire packet size, followed optionally by a specialised adjustment for ATM or PTM framing. This algorithm is shown in Algorithm 1.

Using the network-layer packet size and adding a manually configured overhead value is required because the values reported by the kernel are often wrong due to idiosyncrasies of the CPE unit. While this does make configuration a bit more complex, we seek to alleviate this by providing keywords for commonly used configurations.

As part of the overhead compensation, CAKE also optionally splits "super packets" generated by hardware offload features. These super packets are essential for operating at high bandwidths, as they help amortise fixed network stack costs over several packets. However, at lower bandwidths they can hurt latency, in particular when a link with a high physical bandwidth is shaped to a lower rate. For this reason, we conditionally split super packets when shaping at rates lower than 1 Gbps. This allows CAKE to ensure low latency at lower rates, while still scaling to full line rate on a 40Gbps link.

B. Flow Isolation and Hashing

CAKE replaces the direct hash function used in FQ-CoDel with an 8-way set-associative hash. While set-associative hashing has been well-known for decades as a means to improve the performance of CPU caches [22], it has not seen much use in packet scheduling. Conceptually, a k -way set-associative hash with n total buckets can be thought of as a plain hash with n/k buckets that is only considered to have a collision if more than k items hash into the same bucket. As can be seen in Figure 1, this significantly reduces the hash collision probability up to the point where the number of flows is larger than the number of queues.²

1) *Host Isolation:* With flow fairness, hosts can increase their share of the available bandwidth by splitting their traffic

Algorithm 1 Shaping and overhead compensation algorithm. T_{next} is the time at which the next packet is eligible for transmission.

```

1: function ENQUEUE( $pkt$ )
2:    $net\_len \leftarrow pkt.len - NETWORK\_OFFSET(pkt)$ 
3:    $adj\_len \leftarrow net\_len + overhead$ 
4:   if ATM framing is enabled then
5:      $adj\_len \leftarrow CEILING(adj\_len / 48) * 53$ 
6:   else if PTM framing is enabled then
7:      $adj\_len \leftarrow CEILING(adj\_len / 64) * 65$ 
8:    $pkt.adj\_len \leftarrow adj\_len$ 
9:   if backlog is zero and  $T_{next}$  is after Now then
10:     $T_{next} \leftarrow Now$ 
11: function DEQUEUE
12:   if  $T_{next}$  is after Now then
13:     Schedule interrupt at  $T_{next}$ 
14:   return Nil
15:    $pkt \leftarrow$  Choose Packet
16:    $T_{next} \leftarrow T_{next} + pkt.adj\_len * time\_per\_byte$ 
17:   return  $pkt$ 

```

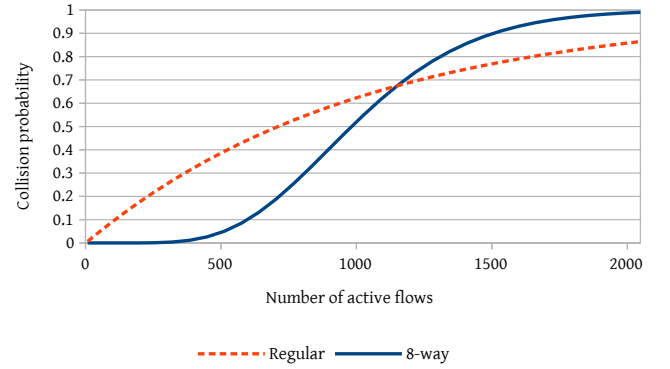


Fig. 1: Probability that a new flow will experience a hash collision, as a function of the number of active flows. 1024 total queues.

over multiple flows. This can be prevented by providing host fairness at the endpoint IP address level, which CAKE can do in addition to flow fairness.

The host isolation is simple in concept: The effective DRR quantum is divided by the number of flows active for the flow endpoint. This mechanism can be activated in three different modes: source address fairness, in which hosts on the local LAN receive equal share, destination address fairness, in which servers on the public internet receive an equal share, or "triple isolate" mode, in which the maximum of the source and destination scaling is applied to each flow. CAKE also hooks into the Linux kernel Network Address Translation (NAT) subsystem to obtain the internal host address of a packet, which would otherwise be obscured since packets are queued after NAT is applied.

CAKE accomplishes this scaling as shown in Algorithm 2: When a packet is enqueued it is hashed into a queue using the transport layer port numbers along with the source and destination IP addresses. In addition, two separate hashes are performed on the packet destination IP address and source IP address. A separate set of hash buckets is kept for these address

²See how we computed these probabilities in the online appendix.

Algorithm 2 Host isolation algorithm.

```
1: function ENQUEUE(pkt)
2:   flow_hash ← HASH(pkt.hdr)
3:   src_hash ← HASH(pkt.src_ip)
4:   dst_hash ← HASH(pkt.dst_ip)
5:   flow ← flows[flow_hash]
6:   if flow is not active then
7:     hosts[src_hash].refcnt_src++
8:     hosts[dst_hash].refcnt_dst++
9:     flow.active ← 1
10:    flow.src_id ← src_hash
11:    flow.dst_id ← dst_hash
12: function GET_QUANTUM(flow)
13:   refcnt_src ← hosts[flow.src_id].refcnt_src
14:   refcnt_dst ← hosts[flow.dst_id].refcnt_dst
15:   host_load ← MAX(refcnt_src, refcnt_dst, 1)
16:   return flow.quantum/host_load
```

hashes. These buckets do not contain a queue of packets, but instead a data structure that keeps two reference counts for each IP address, which track the number of active flows with the given address as source and destination, respectively.

The per-IP reference counts are used to modify the quantum for each active flow. When a flow is scheduled, its "host load" is calculated as the maximum of the reference counts for its source and destination IP addresses. The effective quantum of the flow is simply divided by this load value, which achieves the desired scaling.

C. DiffServ handling

CAKE provides a small number of preset configurations, which map each DiffServ code point into a priority tier. If the shaper is in use, each priority tier gets its own virtual clock, which limits that tier's rate to a fraction of the overall shaped rate. When dequeuing a packet, the algorithm simply picks the highest-priority tier which both has queued traffic and whose schedule is due, if one exists. To allow tiers to borrow excess bandwidth from one another, the dequeue algorithm also tracks the earliest schedule time of all non-empty tiers, and if no other eligible tier is available, that tier is picked instead (within the overall shaper limits).

When the shaper is not in use, CAKE instead uses a simple weighted DRR mechanism to schedule the different priority tiers, with the same weights as the shaper fractions mentioned above. This has weaker precedence guarantees for high-priority traffic, but provides the same proportional capacity reservation and the ability to borrow spare capacity from less than fully loaded tiers.

CAKE defaults to a simple, three-tier mode that interprets most code points as "best effort", but places CS1 traffic into a low-priority "bulk" tier which is assigned 1/16 of the total rate, and a few code points indicating latency-sensitive or control traffic (specifically TOS4, VA, EF, CS6, CS7) into a "latency sensitive" high-priority tier, which is assigned 1/4 rate. The other DiffServ modes supported by CAKE are a 4-tier mode matching the 802.11e precedence rules [17], as well as two

8-tier modes, one of which implements strict precedence of the eight priority levels.

D. ACK filtering

CAKE contains an ACK filtering mechanism that drops *redundant* ACKs from a TCP flow. The mechanism takes advantage of the per-flow queueing by scanning the queue after every packet enqueue, to identify a pure ACK (i.e., an ACK with no data) that was made redundant by the newly enqueued packet. An ACK is only filtered if the newly enqueued packet contains an acknowledgement of *strictly more* bytes than the one being filtered. In particular, this means that duplicate ACKs are not filtered, so TCP's fast retransmit mechanism is not affected. In addition, the filter parses TCP headers and only drops a packet if that will not result in loss of information at the sender; and packets with unknown headers are never dropped, to avoid breaking future TCP extensions. The filter has two modes of operation: a conservative mode that will always keep at least two redundant ACKs queued, and an aggressive mode, that only keeps the most recently enqueued ACK.

IV. PERFORMANCE EVALUATION

In this section, we present a performance evaluation of CAKE. All tests are performed on a testbed that emulates a pair of hosts communicating through a low-bandwidth link. We use the Flent testing tool [23] to run the tests, and the data files are available on the companion web site.¹ Unless otherwise stated below, all tests are run on a symmetrical 10 Mbps link with 50 ms baseline latency. Our basic test is the Real-Time Response Under Load test, which consists of running four TCP flows in each traffic direction, along with three different latency measurement flows [24].

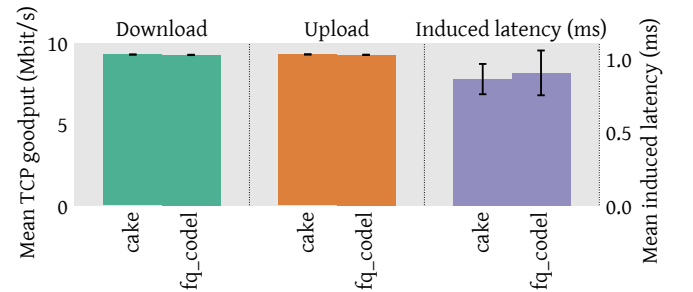


Fig. 2: Baseline throughput and latency of CAKE and FQ-CoDel on a 10 Mbps link.

As can be seen in Figure 2, the baseline performance of CAKE is comparable to that of FQ-CoDel: both achieve low latency and high throughput in the baseline test. This is expected, since CAKE is derived from FQ-CoDel. For a more comprehensive comparison of FQ-CoDel with other queue management algorithms, we refer the reader to [1]. Instead, the remainder of this evaluation focuses on the features outlined in the previous sections.

A. Host Isolation

To evaluate the host isolation feature of CAKE, we run a varying number of TCP flows between two source hosts and four destination hosts. Source host A runs one flow to each of destination hosts A and B, and two flows to destination host C, while source host B runs one flow to each of destination hosts C and D. This makes it possible to demonstrate the various working modes of CAKE’s host isolation feature.

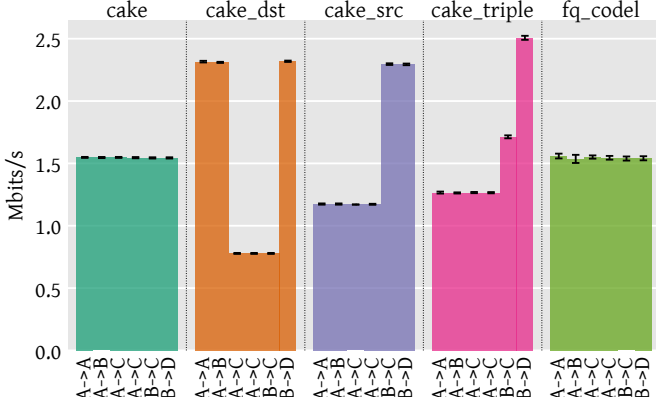


Fig. 3: Host isolation performance with TCP flows from two source hosts to four destination hosts. The columns show different algorithms; each bar shows the average flow goodput.

The result of this test is shown in Figure 3. It shows four configurations of CAKE (no host isolation, source host isolation, destination host isolation and triple isolation) and a test with FQ-CoDel as the queue management algorithm. As can be seen in the figure, both FQ-CoDel and CAKE with no host isolation provide complete fairness between all six flows.

The figure also clearly shows the various modes of flow isolation supported by CAKE: In destination fairness mode (second column), the four destination hosts get the same total share, which results in each of the three flows to destination host C getting 1/3 of the bandwidth of the three other hosts (which only have one flow each). Similarly, in source fairness mode (third column), the two source hosts share the available capacity, which results in the two flows from source B getting twice the share each compared to the four flows from host A.

In the triple isolation case, we see the flow bandwidths correspond to the quantum scaling outlined in Algorithm 2: The first four flows get their quantum scaled by 1/4 since there are four flows active from host A. The fifth flow gets its quantum scaled by 1/3 since there are three flows active to host C. And finally, the last flow gets its quantum scaled by 1/2 as there are two flows active from host B.

B. DiffServ Handling

To demonstrate the DiffServ prioritisation features of CAKE we perform two tests: An RRUL test with each flow marked with a different DiffServ priority, and another test where a high-priority fixed-rate flow competes with several TCP flows.

The result of the former test is seen in Figure 4. This shows that when DiffServ mode is not enabled, all four flows

get the same share of the available bandwidth, while in the DiffServ-enabled case, the Best Effort (BE) flow gets most of the bandwidth. This latter effect is important for two reasons: First, it shows that a flow marked as background (BK) is successfully de-prioritised and gets less bandwidth. Secondly, it shows that the high-priority flows (CS5 and EF) are limited so as to not use more than the share of the bandwidth allocated to the high-priority DiffServ classes.

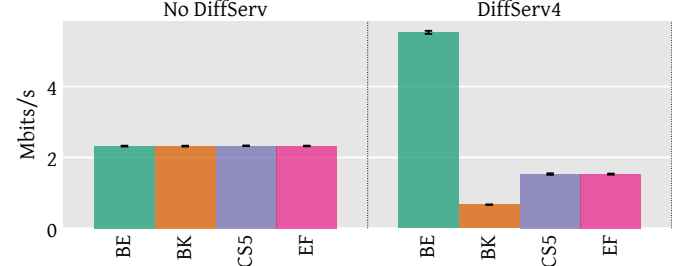


Fig. 4: TCP flows on different DiffServ code points.

To look at the latency performance of a high-priority flow, we turn to Figure 5. This shows the latency over time of a fixed-rate 2 Mbps flow, which marks its packets with the high-priority EF DiffServ marking. This is meant to represent a real-time video conversation. In the test, the flow competes with 32 bulk TCP flows. As can be seen in the figure, both FQ-CoDel and CAKE with DiffServ prioritisation disabled fail to ensure low latency for the high-priority flow. Instead, when the bulk flows start after five seconds, a large latency spike is seen, since the real-time flow has to wait for the initial packets of the 32 TCP flows. This causes the real-time flow to build a large queue for itself (since it does not respond to congestion signals), which then drains slowly back to a steady state around 200 ms (for CAKE) or oscillating between 50 and 500 ms (for FQ-CoDel). In contrast, the DiffServ-enabled CAKE keeps the real-time flow completely isolated from the bulk TCP flows, ensuring it sees no added latency over the duration of the test.

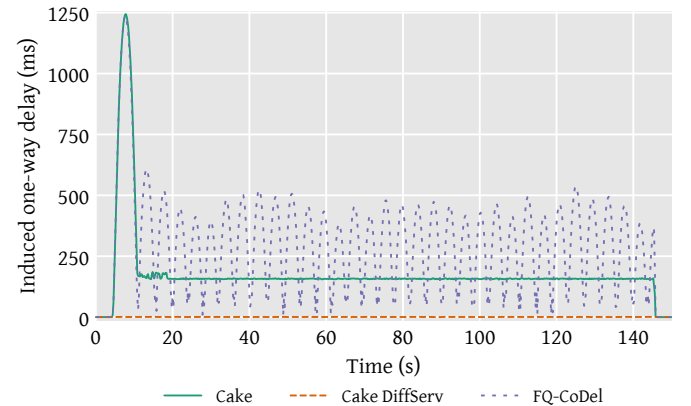


Fig. 5: Latency over time of a 2 Mbps fixed-rate flow with 32 competing bulk flows on a 10 Mbps link. The Y-axis shows additional latency above the base latency of 50 ms. The bulk flows start after 5 seconds.

C. ACK Filtering

Figure 6 shows the performance of ACK filtering on a highly asymmetrical link with 30 Mbps download capacity and only 1 Mbps upload capacity. On this link, we run four simultaneous TCP uploads and four simultaneous TCP downloads. The results of this are shown in Figure 6, which shows the aggregate throughput of all four flows in each direction, along with the added latency of a separate measurement flow. Values are normalised to the baseline without ACK filtering to be able to fit on a single graph. As the figure shows, we see a goodput improvement of around 15% in the downstream direction caused by either type of ACK filtering, which shows that insufficient bandwidth for ACKs can impact transfers in the other direction. For upload, the conservative filtering increases goodput by about 10%, while the aggressive filtering increases throughput by as much as 40%, simply by reducing the bandwidth taken up by ACK packets. We attribute the increase in latency to increased congestion in the downlink direction, which is alleviated somewhat by fewer ACKs being queued in the upstream direction in the aggressive case. The absolute magnitude of the latency increase is only 5 ms.

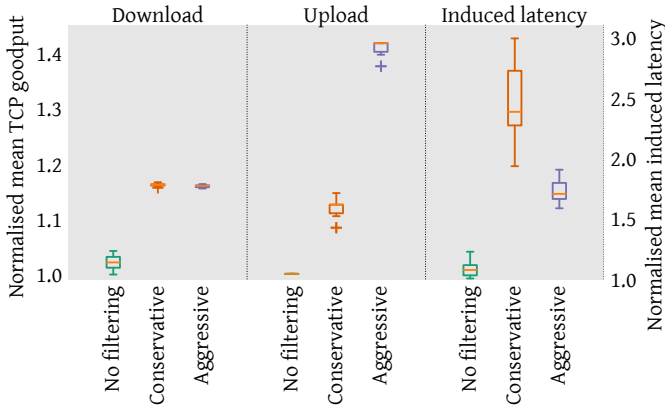


Fig. 6: ACK filtering performance on a 30/1 Mbps link. The graph scales are normalised to the "No filtering" case. The download and upload value ranges are 24.5-27.5 Mbps and 0.45-0.7 Mbps, respectively. The latency range is 2.6-7.5 ms.

V. CONCLUSIONS

CAKE is a comprehensive queue management system for home gateways, that packs several compelling features into an integrated solution, with reasonable defaults to ease configuration. These features include: bandwidth shaping with overhead compensation for various link layers; reasonable DiffServ handling; improved flow hashing with both per-flow and per-host queueing fairness; and filtering of TCP ACKs.

Our evaluation shows that these features offer compelling advantages, and we believe CAKE has the potential to significantly improve the performance of last-mile internet connections. CAKE is open source and ready for deployment, and already ships in the OpenWrt router firmware distribution.

ACKNOWLEDGEMENTS

The authors would like to thank the Bufferbloat and OpenWrt communities for their work on the implementation and testing of CAKE. In particular, Kevin Darbyshire-Bryant was instrumental in enabling NAT-awareness, Ryan Mounce contributed the original ACK filtering code, Sebastian Moeller helped get the overhead compensation right and Anil Agarwal helped with the hash collision probability calculations.

REFERENCES

- [1] T. Høiland-Jørgensen, P. Hurtig, and A. Brunstrom, "The Good, the Bad and the WiFi: Modern AQMs in a Residential Setting," *Computer Networks*, vol. 89, pp. 90 – 106, 2015.
- [2] I. Järvinen and M. Kojo, "Evaluating CoDel, PIE, and HRED AQM techniques with load transients." IEEE, 2014.
- [3] N. Khademi, D. Ros, and M. Welzl, "The new AQM kids on the block: Much ado about nothing?" *Technical Report, Oslo University*, 2013.
- [4] T. Høiland-Jørgensen, D. Täht, and J. Morton, "Piece of CAKE: A Comprehensive Queue Management Solution for Home Gateways," Apr. 2018. [Online]. Available: <https://doi.org/10.5281/zenodo.1226887>
- [5] J. Corbet, "Network transmit queue limits," LWN Article, August 2011. [Online]. Available: <https://lwn.net/Articles/454390/>
- [6] T. Høiland-Jørgensen, M. Kazior, D. Täht, P. Hurtig, and A. Brunstrom, "Ending the Anomaly: Achieving Low Latency and Airtime Fairness in WiFi," in *2017 USENIX Annual Technical Conference (USENIX ATC 17)*. Santa Clara, CA: USENIX Association, 2017, pp. 139–151.
- [7] G. Niestegge, "The 'leaky bucket' policing method in the atm (asynchronous transfer mode) network," *International Journal of Communication Systems*, vol. 3, no. 2, pp. 187–197, 1990.
- [8] A. Eleftheriadis and D. Anastassiou, "Constrained and general dynamic rate shaping of compressed digital video," in *International Conference on Image Processing*, vol. 3. IEEE, 1995, pp. 396–399.
- [9] V. P. Rao, M. P. Tahiliani, and U. K. K. Shenoy, "Analysis of sfqCoDel for active queue management," in *ICADIWT 2014*. IEEE, 2014.
- [10] R. Adams, "Active queue management: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1425–1476, 2013.
- [11] N. Benameur, F. Guillemin, and L. Muscariello, "Latency reduction in home access gateways with shortest queue first," in *Proc. ISOC Workshop on Reducing Internet Latency*, 2013.
- [12] T. Høiland-Jørgensen, P. McKenney, D. Täht, J. Gettys, and E. Dumas, "The Flow Queue CoDel Packet Scheduler and Active Queue Management Algorithm," RFC 8290, RFC Editor, Jan. 2018.
- [13] M. Shreedhar and G. Varghese, "Efficient fair queuing using deficit round-robin," *IEEE/ACM Transactions on Networking*, Jun. 1996.
- [14] A. Kortebe, S. Oueslati, and J. Roberts, "Implicit service differentiation using deficit round robin," *ITC19*, 2005.
- [15] S. O. Abdesslem Kortebe, "Cross-Protect: Implicit Service Differentiation and Admission Control," pp. 56 – 60, 2004.
- [16] J. Babiarz, K. Chan, and F. Baker, "Configuration Guidelines for DiffServ Service Classes," RFC 4594 (Informational), RFC Editor, Aug. 2006, updated by RFC 5865.
- [17] T. Szigeti, J. Henry, and F. Baker, "Mapping Diffserv to IEEE 802.11," RFC 8325 (Proposed Standard), RFC Editor, Feb. 2018.
- [18] W.-S. Hwang and P.-C. Tseng, "A qos-aware residential gateway with bandwidth management," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 3, pp. 840–848, 2005.
- [19] H. Wu, J. Wu, S. Cheng, and J. Ma, "Ack filtering on bandwidth asymmetry networks," in *APCC/OECC'99*, vol. 1. IEEE, 1999.
- [20] L. Storf, "Enhancing cable modem tcp performance," *Texas Instruments Inc. white paper*, 2003.
- [21] H. Kim, H. Lee, S. Shin, and I. Kang, "On the cross-layer impact of tcp ack thinning on ieee 802.11 wireless mac dynamics," in *64th Vehicular Technology Conference*. IEEE, 2006.
- [22] A. J. Smith, "A comparative study of set associative memory mapping algorithms and their use for cache and main memory," *IEEE Transactions on Software Engineering*, no. 2, pp. 121–130, 1978.
- [23] T. Høiland-Jørgensen, C. A. Grazia, P. Hurtig, and A. Brunström, "Flent: The FLExible Network Tester," in *VALUETOOLS*, 2017.
- [24] D. Täht, "Realtime response under load (rrul) test," November 2012. [Online]. Available: https://www.bufferbloat.net/projects/bloat/wiki/RRUL_Spec/