

General comments:

- SLA → SLO (find, replace)
- Propagate other comments across the text.
- Page numbers!
- More descriptive Figure captions; by reading the Figure caption I should be able to tell which workload is used; and what type of experiment it is
- Workload 2 is not a very low latency search; we can say it's any low latency workload, e.g. in-memory database.

Also it is not imitating the Facebook workload quite, but the service time is in the order of mag. for an in-memory database.

Implementation and end-to-end Evaluation of the HULL Architecture

(&&) due to large amount of queried data that needs to be served from memory of many servers.

(**), thus, latency variability needs to be bounded. One of the sources of latency variability are queuing delays that need to be minimized

School of Computer & Communication Sciences, EPFL, Switzerland

Semester Project

Prasopoulos Konstantinos

applications, such as Web search, online store etc.,

some amount of queuing is necessary to maintain acceptable throughput levels.

Abstract—User-facing data center workloads have very strict tail latency requirements in order to offer a consistent user experience. These workloads often involve the partitioning of queries to multiple servers and the aggregation of the responses. Thus, network-induced delays from queuing in particular need to be minimized. The goal of this project is to implement the HULL architecture and evaluate it on such workloads via simulation. HULL aims for practically no queuing time along with acceptable bandwidth by employing three techniques: congestion signals calculated based on part of the link capacity to avoid queue buildup, packet pacing for throughput-heavy flows to reduce queuing time variability and DCTCP in order to respond to congestion proportionally and avoid throughput loss. In tested workloads with requirements in the order of tens of milliseconds, HULL performs no better than DCTCP and up to 40% better than TCP Reno, depending on the background load. On the other hand, in workloads with tighter requirements (hundreds of microseconds - e.g. in-memory key-value store queries), HULL reduces query completion latency by 32% compared to DCTCP.

TODO: introduce this apps as OLDI applications (On-line Data Intensive), then just use OLDI in the rest of the report

The way users experience the Internet, be it interacting on social media, querying a search engine etc., is highly dependant on fast response times. Upon arriving to a data center, such user requests often follow the partition/aggregate pattern in which the request is ~~fragmented~~ multiplied and assigned to multiple servers and the responses are aggregated to produce the final page content. Satisfying the request either means that all the servers must respond or it involves returning a partial, lower quality result. This decision is made based on response deadlines called **Service Level Agreements** (SLAs). Emphasis is given on the 99th percentile of overall response times in order to provide a consistent experience for ~~all~~ the vast majority of users. The fact that satisfying the overall request depends on the slowest server means that it is very hard to keep tight SLOs tail latencies, especially as the number of servers increases (The Tail at Scale problem [1]).

Partition/aggregate workloads share the data center network infrastructure with throughput-sensitive, large background flows that synchronize the state of the servers. TCP can provide good throughput for the background flows but, because it detects congestion only after network buffers are full, it significantly reduces the performance of latency-sensitive flows. A full switch buffer can add several milliseconds of latency which will then not be available for additional processing. It is thus important to operate with limited buffering within the network. At the same time,

Not just to sync servers; this flows can e.g. belong to batch applications (some map-reduce jobs) that are best-effort applications which are not latency sensitive

A different kind of congestion control that combines low network-induced latency through minimal queuing and acceptable throughput is necessary. The HULL (High-bandwidth Ultra-Low Latency) architecture [2] attempts this by employing three mechanisms. First, it uses *Phantom Queues* (PQs). PQs are positioned on a link and behave like queues that drain at a lower rate than the link's capacity (without actually queuing packets). In this way, this virtual queue will build-up and signal congestion earlier than a switch buffer draining at the nominal rate. Therefore, the PQ attempts to keep the aggregate rate of the senders below that which the switch can handle in order to achieve minimal buffering. PQs use ECN (Explicit Congestion Notification) marking in conjunction with the second mechanism, DCTCP (Data Center TCP) [3]. In contrast to TCP which reacts to the presence of congestion by halving the congestion window, DCTCP reacts to the extent of congestion (estimated by the number of ECN-marked packets) in a proportional manner. DCTCP also utilizes the active queuing capabilities of switches in order to maintain queue occupancies near the ECN marking threshold. Finally, packet pacing is employed for large flows in order to reduce queuing spikes that would widen the variance of queuing delay and would also limit the throughput by increasing the number of packets the PQ marks.

The objective of this project is the implementation of the HULL architecture on ns-2 (Network Simulator 2) [4] and its end-to-end evaluation with partition/aggregate, user-facing workloads which the original paper [2] does not cover. The simulation results show that:

- In workloads with **SLAs** of tens of milliseconds, HULL is a significant improvement over TCP (~~in the order of ~20% lower WHAT?~~ in ~~metric? 99th pct latency?~~) and is equivalent to DCTCP.
- In very low latency workloads with **SLAs** of hundreds of microseconds, HULL outperforms TCP by a factor of 6 and DCTCP by close to 32%. ~~metric? 99th pct latency?~~ again, what is the metric?
- Overall, HULL manages to keep queuing latencies in the order of 10 microseconds while DCTCP achieves an order of magnitude higher. Therefore, only very low latency applications can benefit from the 100 microseconds ~~or so~~ that HULL saves.
- As the network load increases, the performance of HULL with partition/aggregate workloads decreases

Unclear;
what do you mean by additional processing, and what won't be available?

~~converges~~
and tends towards that of DCTCP.

II. SYSTEM DESIGN AND IMPLEMENTATION

This section describes the three components of the HULL architecture and their implementation ~~in~~ for ns-2.

A. Phantom Queues

Phantom queues create bandwidth headroom in order to avoid the need for buffering headroom. The goal is to signal congestion before any queuing takes place. This is in contrast to TCP that relies on packets being dropped by a full queue and to DCTCP that uses a queue occupancy threshold for ECN marking. In the HULL architecture, PQs behave as normal queues that drain at a lower rate than the link's capacity and are placed in series with the egress ports of switches. They do not actually store packets; Instead they maintain a counter that is increased upon a new packet arrival and decreased according to the drain rate. If the counter exceeds a threshold, the CE (congestion experienced) IP header field of the packet responsible is set. In this way, the PQ controls the aggregate rate of the senders so that it is below that of the in-series switch port.

The two parameters (drain rate and marking threshold) are tuned in the original paper [2] such that the PQ reacts to congestion quickly and avoids build-up at the switch buffer. The drain rate does not need to be set low to achieve practically zero queuing delay (a baseline of 95% of the link rate is used). In order to facilitate a quick response to congestion, the baseline setting for the marking threshold is 1KB. The two parameters are inversely related. As the drain rate increases, the virtual queue counter takes longer to reach the threshold and therefore a lower threshold value is needed in order to react to congestion quickly.

For ns-2, the phantom queue was implemented as a component of the SimpleLink class that unidirectionally connects two nodes and simulates packet queuing, transmission and propagation (a switch node does not simulate queuing itself - instead its outgoing links have queues that represent the port buffer). The PQ was placed after the queue object and all dequeued packets go through it. Making sure that only egress switch links have PQs is done at ns-2's OTcl (Object oriented Tcl) [5] front-end. This asymmetry means that two different opposing simplex links (~~the~~ the OTcl SimpleLink objects) are used to connect a switch and a host instead of a duplex link (a duplex link is just a wrapper that creates two identical simplex ones).

B. DCTCP

Using ECN-capable TCP in conjunction with phantom queues would lead to significant loss of throughput because TCP halves its congestion window when congestion is detected - which ~~happens~~ is very often when buffering is meant to be minimal—and thus there is no elasticity to absorb bursts of traffic. For this reason, DCTCP is used instead

~~and it~~
which operates as follows: The DCTCP sender estimates the amount of congestion on the flow path based on the ratio of ACKs (acknowledgments) that have the ECN-Echo (ECE) field set. Normally, it is an ECN-capable switch that sets the congestion experienced (CE) IP header field of packets that cause the queue length to exceed a threshold. With HULL, the task of setting CE is removed from the switch and assigned to the PQ. In both cases, the DCTCP receiver appropriately sets the ECE TCP header field of the ACKs it sends so that the sender can know how many of the packets it sent were marked. Finally, the sender adjusts its window in proportion to the estimate of congestion it maintains thus avoiding the sawtooth behavior of TCP. The DCTCP algorithm maintains the queue length close to the marking threshold because the correction it does on the window of the sender depends on ~~by~~ how much this threshold is violated.

A DCTCP patch for the FullTcp agent of ns-2 is available (see appendix). FullTcp needs to be configured in the OTcl front-end so that it implements the DCTCP algorithm (examples of this can be found in this project's repository - see appendix).

C. Pacing

The use of phantom queues with such small marking thresholds suggests that any flow that sends more than a few packets back-to-back in a burst will receive multiple congestion signals. Additionally, the burst of traffic will temporarily increase the queuing latency at the switch before the sender gets congestion feedback. This will increase the tail latencies of concurrent small workloads. Therefore, in order to reduce queuing spikes and to maintain acceptable throughput for large flows, HULL uses pacing at the hosts to add temporal spacing between the packets of a burst. Since pacing introduces latency, only non-latency-sensitive, throughput-heavy flows are paced.

The HULL pacer is placed after a host's NIC (~~Network Interface Controller~~) and has two main tasks. The first is to measure the momentary sending rate of the host and slowly adapt its own rate according to the following update rule:

$$R_{tb} \leftarrow (1 - \eta) * R_{tb} + \eta * M_r / T_r + \beta * Q_{tb}$$

~~where~~

R_{tb} is the sending rate, M_r is the number of bytes received in the last T_r seconds, Q_{tb} is the number of bytes that the pacer has in its queue and η , β are positive constants. The pacer updates the rate every T_r seconds. The term $\beta * Q_{tb}$ is needed to increase the rate as the queue length increases.

The second task is to make sure that congestion-causing large flows are paced and small flows are not. To this end, the pacer maintains a flow association table. Initially all flows are not associated. When the pacer observes an ACK with the ECE field set, it associates its flow with probability p_a (to avoid pacing small flows). The association times out after T_i seconds.

For ns-2, the HULL pacer was again implemented as a component of the SimpleLink class and placed after the queue (represents the NIC of the host in this case) and before the transmission/propagation component. Adding a pacer to a link is again performed in OTcl.

The implementation is as close as possible to the original hardware implementation of a token bucket rate limiter. Whenever a packet arrives from the host, the pacer checks if it belongs to an associated flow. If not, the packet is forwarded with no delay. Otherwise, the pacer checks if there is another packet already queued up in which case it enqueues this packet as well (FIFO). If there are no packets in the queue, the pacer calculates whether there are enough tokens available. If there are, the packet is forwarded and its size is subtracted from the number of available tokens. If not, the packet is queued up and its departure is scheduled based on the current token acquisition rate (R_{tb}). When this scheduled event happens, the next packet in the queue, if it exists, is scheduled similarly.

Network	Link Speed = 1Gbps, Link Latency = 5μs, Port Queue = 500pkts
TCP	Max. Window = 1256, Max. ssthresh. = inf , MTU = 950bytes, Min. RTO = 10ms, Nagle: Disabled
DCTCP	$g = 1/16$, K = 30KB
Pacer	Bucket Depth = 3KB, $\eta = 0.125$, $\beta = 16$, $p_a = 0.125$, $T_i = 10ms$
Phantom Queue	Drain Rate = 95%, Threshold = 1KB

Table I: General System Parameters

There are also two asynchronous loops: one that updates the token level (up to a configurable maximum) based on the current rate and one which updates the current rate (R_{tb}) every T_r seconds. When the rate update occurs, the departure time of the packet at the head of the queue is recalculated in order to avoid scenarios in which a packet was scheduled long into the future based on a near zero rate.

Finally, because the pacer sits on the unidirectional link that connects the host to a switch, it does not have access to the ACKs which flow in the opposite direction. This problem is circumvented by the addition of a "gotecnecho" field in the TCP header (already added to FullTcp by Mohammad Alizadeh based on the code comments - probably in [this citation](#) to the link where you found code). Outgoing packets use this field to inform the pacer that ECE set ACKs were received. When such a packet is received by the pacer, it associates its flow with probability p_a . If the flow is associated, a disassociation event is scheduled after T_i seconds. This event is rescheduled if the flow gets associated again within T_i seconds.

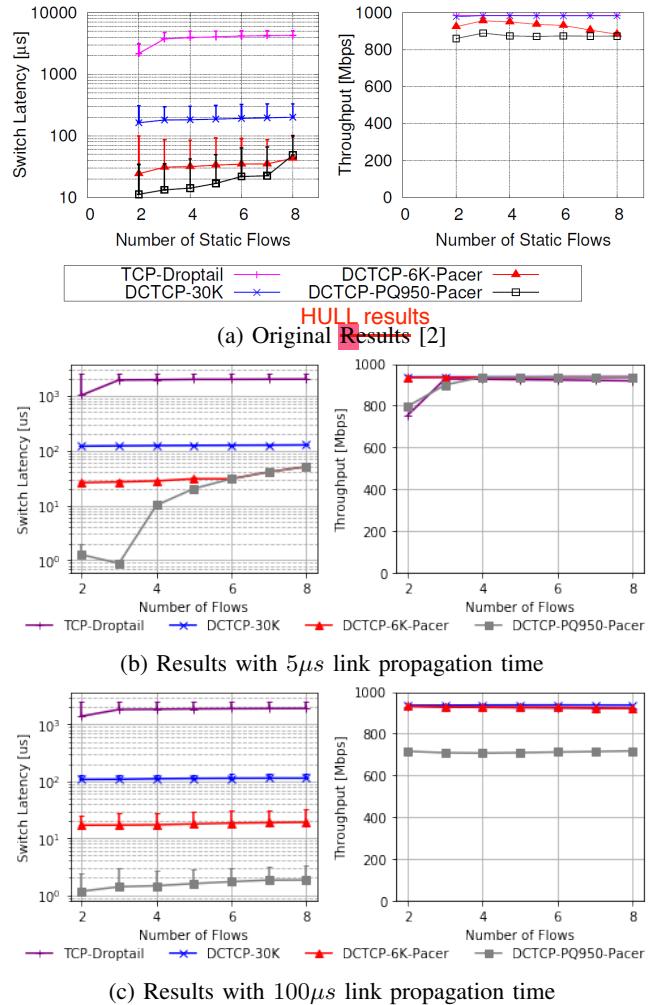


Figure 1: Static Flow Experiments. Left: Queuing latency at the congested client switch port (log scale). Right: Aggregate throughput of all flows. The error bars represent the 99th percentile.

III. RESULTS

This section presents the results obtained by simulating the HULL architecture under a variety of workloads using ns-2. First, this implementation is compared to the experimental findings of the original paper. Second, the performance of HULL is evaluated under two types of partition/aggregate workloads. The architecture is compared to TCP Reno and DCTCP (with the recommended parameters and in a more aggressive configuration). The network topology simulates a single rack with a configurable number of servers connected to the TOR (Top of Rack) switch. In all simulations, one of the servers behaves as a client and issues requests to the other servers. The responses cause congestion at the client's switch port. Table I summarizes the system parameters which hold unless specified otherwise. The MTU is set to 950Bytes so that the phantom queues configured

		Switch Latency (μs)		900B FCT (μs)		10MB FCT (ms)	
		Avg	99th	Avg	99th	Avg	99th
20% Load	TCP-Drop-tail	86.2	2284.4	141.7	2387.5	101.8	258.9
	DCTCP-30K	7.1	133.1	59.7	792.0	103.1	276.4
	DCTCP-6K-Pacer	4.2	124.2	59.9	800.1	105.0	275.4
	DCTCP-PQ950-Pacer	0.3	9.7	40.3	74.7	200.5	304.1
40% Load	TCP-Drop-tail	352.3	3412.3	430.3	3560.1	134.2	447.8
	DCTCP-30K	26.9	204.5	87.6	913.0	139.2	427.6
	DCTCP-6K-Pacer	15.2	170.0	84.3	908.1	144.7	429.9
	DCTCP-PQ950-Pacer	2.8	57.8	45.1	128.6	234.9	481.4
60% Load	TCP-Drop-tail	811.7	3718.0	869.3	3791.5	206.2	824.2
	DCTCP-30K	57.6	232.2	117.1	843.5	219.2	729.4
	DCTCP-6K-Pacer	30.0	188.5	103.2	846.5	228.6	732.0
	DCTCP-PQ950-Pacer	12.9	106.1	55.4	150.5	313.3	749.2

Table II: Dynamic flow simulation results. The displayed load level refers to the traffic level on the client-switch link.

with a $1KB$ threshold (as in the original paper) do not mark all maximum sized packets even if there is no other traffic (the PQ increments the virtual queue length upon receiving a packet as described in 5.1 of [2]). The retransmission timeout was set to $10ms$ according to the DCTCP paper [3]. The port buffer's fixed size was selected to be similar to the dynamically allocated sizes of the buffers in the original HULL paper.

A. Implementation Evaluation

In order to make sure that this implementation of HULL behaves as intended, two of the experiments of the original paper were reproduced. The results were not expected to be identical as there is a wide range of differences between the original experimental setup and this simulation. Some of those differences are: 1) ns-2 uses fixed size switch buffers compared to the shared, dynamically allocated memory of the Broadcom Triumph2 switch used in the paper, 2) software overhead or features like Interrupt Coalescing are not simulated at the servers, 3) the exact details of the original physical system are not known (e.g. propagation delay, TCP configuration etc.). Nevertheless, the fundamental properties are the same and HULL generally manages to keep the congested port's buffer mostly empty by sacrificing some of the throughput of the large flows.

Static Flow Experiments: This experiment aims to measure throughput and the congested port's queue occupancy under heavy and continuous load. To this end, a client sends requests for the same file to a static number of servers which in turn respond by sending data as fast as possible. The congested port's queue length is sampled every 1ms and is used to calculate the resulting queuing latency. As in the original paper, four schemes are compared: (1) TCP with drop-tail buffers, (2) DCTCP with a $30KB$ marking threshold, (3) DCTCP with a $6KB$ marking threshold along with pacing and (4) DCTCP with both pacers and phantom queues with the marking threshold at $1KB$ and the drain rate at $900Mbps$ (the original paper uses a $950Mbps$ drain rate but it was quite ineffective in this experiment). In scheme 4, it is the PQs and not the switch that are responsible for

setting the packets' CE field.

The simulation results can be seen in Figure 1b. In general, HULL reduces the switch latency (i.e. the queue occupancy) by two orders of magnitude compared to TCP and by one compared to DCTCP as is the case in the original results (Figure 1a). The present results differ from the original in the following ways: 1) The complete scheme (DCTCP-PQ950-Pacer) converges to the DCTCP-6K-Pacer at lower flow counts, (2) the latency of the complete scheme is higher with 2 flows than with 3 and (3) the complete scheme has the same throughput as the rest for flows 4 to 8.

The three differences are all caused by the way the $5\mu s$ per link propagation delay used in the simulation affects the round trip time (RTT). Given that the sending rate of TCP-like protocols is approximately $congestion\ window/RTT$ and the RTT is about $40\mu s$ ($20\mu s$ to transmit a $1KB$ packet twice and $20\mu s$ for the ACK to return due to propagation delays - not accounting for queuing delay), each additional window length unit adds $MTU/RTT \approx 200Mbps$ to the throughput of the flow. The minimum window size used during the simulation was 2 and at this size each flow contributes a minimum of $400Mbps$. With 3 flows, the total of $1200Mbps$ consistently exceeds the drain rate of the phantom queue between the switch and the client and the sending window is fixed at 2 packets. However, the sender practically ignores the marked ACKs as the sending window is already at its minimum size. With 2 flows present, the average window length is 3.4 packets and the senders are probing for additional bandwidth and reacting to congestion normally (hence the higher latency variation).

Figure 1c further illustrates this point. The experiment is repeated with $100\mu s$ link propagation delay instead of $5\mu s$. It is important to note that this is a far-fetched scenario in which very large flows are synchronized and continuously send data for several seconds. This issue does not affect the rest of the results (at $5\mu s$).

Dynamic Flow Experiments: In this experiment, small $900Byte$ requests are combined with large $10MB$ ones in order to compare how each scheme performs when it

because?



Please add a very simple figure that shows the communication pattern of the partition aggregate

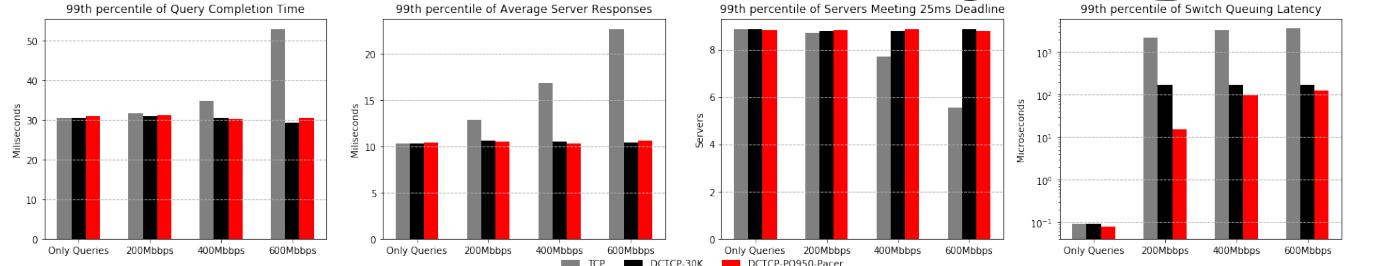


Figure 2: The x-axis varies based on the amount of background traffic along with the partition/aggregate workload (from none to 60% of the link capacity). (a): 99th percentile of query completion times (defined by the slowest server response), (b): 99th percentile of the average server response time (ignores application level variability), (c): 99th percentile of the number of servers that respond within a 20ms deadline (worst cases), (d): 99th percentile of the queuing induced latency at the congested client-switch port buffer.

comes to small flow latency and large flow throughput. In this workload, the client sends requests for small and large files at intervals drawn from two different Poisson distributions configured based on the average target load on the congested switch-client link. For each request, the client asks the servers in a round-robin fashion. The ratio between the number of small and large requests is 4 to 1 while the throughput ratio is 1 to 2,222. The queuing latency at the switch’s port facing the client is calculated based on the momentary queue occupancy. Finally, the flow completion times (FCT) are measured for both the small and large flows.

The simulation results for three levels of link utilization are shown in Table II. The results follow a trend similar to the original paper. Considering the 99th percentile of measurements, the simulation of HULL shows an average switch latency reduction of 98.2% compared to TCP and of 69.6% compared to DCTCP (98.3% and 88.6% in the original paper respectively). Concerning the flow completion times of the small flows, the simulation results indicate an average FCT reduction of 96.4% compared to TCP and of 86.1% compared to DCTCP (89.1% and 60.2% in the original paper). Finally, the FCT of the large flows increased by 0.1% compared to TCP and by 7% compared to DCTCP (20.3% and 35.8% in the original paper). However, the mean FCT for the large flows increases by 62.2% based on the simulation but only by 34.8% based on the original paper’s results.

Given that the simulation results appear to trade-off bandwidth for latency compared to the original paper, it is reasonable to assume that HULL could be configured less aggressively to account for the differences between the simulation and the physical environment. For example, by setting the PQ marking threshold to 1500Bytes for the 20% workload the switch latency increases from 0.3 to 1.4 μ s and from 9.7 to 60.1 μ s for the mean and the 99th percentile respectively. At the same time, the equivalent changes for the small flows are from 40.3 μ s to 46.5 μ s and from 74.7 μ s to 362.2 μ s and for the large flows the FCTs changed from

200.5 μ s to 132.7 μ s and from 304.1 μ s to 289.8 μ s. These results are much closer to the ratios observed in the original paper and also illustrate the sensitivity of HULL to the marking threshold.

B. Partition/Aggregate Workloads

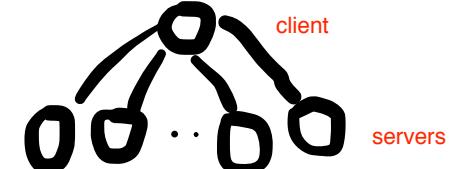
Search: In this section, HULL is evaluated against a web search-like, partition/aggregate workload. The client sends 3500Byte queries at Poisson intervals to 10 servers. Each server processes the query for a duration drawn from:

$$\text{Service time} = 180 * \text{gamma}(0.7, 20000) + 10000 [ns]$$

where $\text{gamma}(0.7, 20000)$ is the gamma distribution with parameters $\kappa = 0.7$ and $\theta = 20000$. The overall mean service time is 2.53 milliseconds. Each server’s response is 2800Bytes. A query is considered complete when the slowest server responds. Each server processes the queries sequentially (not a significant factor at this query rate - 30% server occupancy). The workload parameters are based on OLDISIM [6].

The search workload was also combined with the background dynamic workload described in section III-A in order to simulate a more complete data center environment as described in [3]. The simulation was run at four different levels of average background traffic on the congested 1Gbps switch-client link: (1) no traffic, (2) 200Mbps, (3) 400Mbps, (4) 600Mbps. The search workload generated approximately 12,000 queries and an average of 26.5Mbps on the same link.

The results are displayed in Figure 2. Starting from the rightmost chart, when the background traffic is a reasonable 200Mbps, HULL manages to reduce the 99th percentile of queuing latency to 10 μ s - one order of magnitude lower than DCTCP and two lower than TCP. However, this sub-1ms reduction does not translate to a measurable improvement to the 99th percentile of the query completion times (leftmost chart) which hover around 30ms and are dominated by the tail of the servers’ processing time. The same applies



to the 99th percentile of the average server response time which can be seen in the second chart. When it comes to higher levels of background traffic, both DCTCP and HULL outperform TCP on both the 99th percentiles and the means (not shown) by an appreciable margin. The third chart shows the average number of servers which respond within a 25ms (in the worst 1% of queries -worst defined by the number of servers that respond) and thus reflects the quality of the search if such a deadline was used. Both HULL and DCTCP have tight tail latencies and the quality is minimally affected by the background traffic.

It is worth mentioning that the pacer successfully selects which flows need pacing. For example, with 200Mbps of background traffic, the pacer attached to one of the servers decided to associate the search flow 1695 times out of the 48000 search packets (3.5%) and the background flow 12,071 times out of the 260,000 packets (4.6%). This count includes the times a flow was already being paced - in that case the disassociation timer was reset. The difference between the two flows is that the search flow was disassociated 1396 times (82%) while the background flow was only disassociated 94 times (0.8%). Essentially, even if a search query was paced, the next one was unlikely to be so while every background 10MB transfer was paced.

*Extra low latency workload (**)*

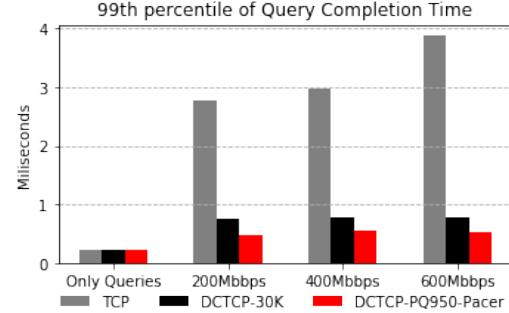
Low latency Search: The conclusion that can be drawn from the previous workload is that while HULL outperforms TCP, the additional 100 μ s of queuing time saved compared to DCTCP need an appropriate, very low latency application, in order to have an impact. Such a partition/aggregate workload is simulated in this section. It is similar to the previous search workload but the size of the queries and the responses is now 100Bytes and critically, the per-server service time is 100 μ s. Approximately 50000 queries are generated along with an average of 4Mbps on the congested client-switch link. *These parameters approximate an in-memory based search workload* (e.g. [7]). *This is not quite true, remove the reference*

The results are presented in Figure 3. Across the scenarios that involve background traffic, HULL reduces the 99th percentile of the query completion time by 84% compared to TCP and by 32% compared to DCTCP. Additionally, HULL offers perfect response quality with a strict 1ms response deadline.

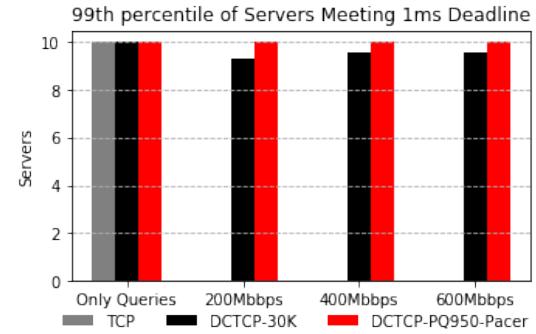
*Extra low latency workload: clarify that this represents e.g. and in-memory database (**)*

IV. CONCLUSION

The goal of this project was the implementation and end-to-end evaluation of the HULL architecture. HULL was implemented for ns-2 and the simulation results from experiments similar to those carried out on a real topology by the original paper authors are qualitatively similar. HULL was evaluated against user-facing, partition/aggregate workloads in order to measure the effect it can have on the user experience. HULL offers great tail latency gains compared to TCP when background traffic is present and an improvement



(a) 99th percentile of query completion times (defined by the slowest server response).



(b) 99th percentile of the number of servers that respond within a 1ms deadline (worst cases).

Figure 3: The x-axis varies based on the amount of background traffic along with the partition/aggregate workload (from none to 60% of the link capacity).

in the order of 30% compared to DCTCP in ultra-low latency applications (sub-1ms).

REFERENCES

- [1] J. Dean and L. A. Barroso, “The tail at scale,” *Commun. ACM*, vol. 56, no. 2, pp. 74–80, Feb. 2013. [Online]. Available: <http://doi.acm.org/10.1145/2408776.2408794>
- [2] M. Alizadeh, A. Kabbani, T. Edsall, B. Prabhakar, A. Vahdat, and M. Yasuda, “Less is more: Trading a little bandwidth for ultra-low latency in the data center,” in *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI’12. Berkeley, CA, USA: USENIX Association, 2012, pp. 19–19. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2228298.2228324>
- [3] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, “Data center tcp (dctcp),” in *Proceedings of the ACM SIGCOMM 2010 Conference*, ser. SIGCOMM ’10. New York, NY, USA: ACM, 2010, pp. 63–74. [Online]. Available: <http://doi.acm.org/10.1145/1851182.1851192>
- [4] “ns-2,” <https://www.isi.edu/nsnam/ns/>.
- [5] “Otel,” <https://en.wikipedia.org/wiki/Otel>.

- [6] “Oldsim,” <https://cloudplatform.googleblog.com/2015/03/benchmarking-web-search-latencies.html>.
- [7] Y. Xu, E. Fraachtenberg, S. Jiang, and M. Paleczny, “Characterizing facebook’s memecached workload,” *IEEE Internet Computing*, vol. 18, no. 2, pp. 41–49, Mar 2014.

We could have done it with this workload, but it does not say anything about the service time though.

We could use the parameters from the zygos paper and try to imitate Silo, but it is a bit too late now.

APPENDIX

SIMULATING HULL ON NS-2

A. ns-2 and DCTCP

HULL was implemented for ns-2 2.34. Straightforward instructions on installing ns-2 on Ubuntu and adding DCTCP can be found [here](#).

B. Adding HULL to ns-2

The modified ns-2 files can be found in the project’s repository. After installing ns-2.34, updating the environment variables and patching-in DCTCP, do:

- Replace ns-2.34/tcl/lib/ns-default.tcl
- Replace ns-2.34/tcl/lib/ns-lib.tcl
- Replace ns-2.34/tcl/lib/ns-link.tcl
- Replace ns-2.34/link/delay.h and delay.cc
- Add hull-pacer.h and hull-pacer.cc to ns-2.34/adc/
- Add ”adc/hull-pacer.o” to the OBJ_CC variable of ns-2.34/Makefile.in
- In ns-2.34/ do: ./configure, then make clean and then make.

C. Reproducing the results

Note: the simulation result files are available in the repository and can be used to avoid running the simulation scripts which in some cases take a day or two to complete.

Reproducing Figure 1b (static flow):

- 1) Run ”static_flow_main_coord.sh” to run all the relevant experiments (takes 10s of minutes).
- 2) In results/static_flow_main (result folder for this script) run the ”postprocess.ipynb” notebook.

Do similarly for Figure 1c by running ”static_flow_100_coord.sh” and then ”results/static_flow_100/postprocess.ipynb”

Reproducing the data from Table II (dynamic flow):

- 1) Run ”coordinator_dynamic.sh”. Folders named ”results/nofanout_(load_level)_9flows” will be created and populated (takes minutes).
- 2) Run ”results/parse_final_coord_dynamic_wrkld.ipynb”.

Reproducing Figure 2 (search workload):

- 1) Run ”coordinator_search.sh”. Folders named ”results/bkg_(load_level)_fanout_wkld1_30_10flows” and ”results/onlyfanout_wkld1_30_10flows” will be created and populated (takes hours).
- 2) Run ”results/parse_final_coord_fanout.ipynb”.

Reproducing Figure 3 (low latency workload):

- 1) Run ”coordinator_100B.sh”. Folders named ”results/bkg_(load_level)_fanout_wkld0_5_10flows” and ”results/onlyfanout_wkld0_5_10flows” will be created and populated (takes days).

- 2) Run ”results/parse_final_coord_fanout_100B.ipynb”.