usenix
THE ADVANCED
COMPUTING SYSTEMS
ASSOCIATION

A High-Speed Load-Balancer Design with Guaranteed Per-Connection-Consistency

Tom Barbette, Chen Tang, Haoran Yao, Dejan Kostić, Gerald Q. Maguire Jr., Panagiotis Papadimitratos, and Marco Chiesa, *KTH Royal Institute of Technology*

https://www.usenix.org/conference/nsdi20/presentation/barbette

This paper is included in the Proceedings of the
17th USENIX Symposium on Networked Systems Design
and Implementation (NSDI '20)
February 25–27, 2020 • Santa Clara, CA, USA
978-1-939133-13-7

Open access to the Proceedings of the
17th USENIX Symposium on Networked
Systems Design and Implementation
(NSDI '20) is sponsored by
NetApp

A High-Speed Load-Balancer Design with Guaranteed Per-Connection-Consistency

Tom Barbette Chen Tang Haoran Yao Dejan Kostić
Gerald Q. Maguire Jr. Panagiotis Papadimitratos Marco Chiesa
KTH Royal Institute of Technology

Abstract

Large service providers use load balancers to dispatch millions of incoming connections per second towards thousands of servers. There are two basic yet critical requirements for a load balancer: *uniform load distribution* of the incoming connections across the servers and *per-connection-consistency* (PCC), *i.e.*, the ability to map packets belonging to the same connection to the same server even in the presence of changes in the number of active servers and load balancers. Yet, meeting both these requirements at the same time has been an elusive goal. Today's load balancers minimize PCC violations at the price of non-uniform load distribution.

This paper presents CHEETAH, a load balancer that supports uniform load distribution and PCC while being scalable, memory efficient, resilient to clogging attacks, and fast at processing packets. The CHEETAH LB design guarantees PCC for *any* realizable server selection load balancing mechanism and can be deployed in both a stateless and stateful manner, depending on the operational needs. We implemented CHEETAH on both a software and a Tofino-based hardware switch. Our evaluation shows that a stateless version of CHEETAH guarantees PCC, has negligible packet processing overheads, and can support load balancing mechanisms that reduce the flow completion time by a factor of $2-3$x.

1 Introduction

The vast majority of services deployed in a datacenter need load balancers to spread the incoming connection requests over the set of servers running these services. As almost half of the traffic in a datacenter must be handled by a load balancer [41], the inability to uniformly distribute connections across servers has expensive consequences for datacenter and service operators. The most common yet cost-ineffective way of dealing with imbalances and meet stringent Service-Level-Agreements (SLAs) is to over-provision [13].

Existing LBs rely on a simple hash computation of the connection identifier to distribute the incoming traffic among the servers [3, 13, 15, 20, 37, 41, 53]. Recent measurements on Google's production traffic showed that hash-based load balancers may suffer from load imbalances up to 30% [13].

A natural question to ask is why existing load balancers do not rely on more sophisticated load balancing mechanisms, *e.g.*, weighted round robin [51],"power of two choices" [33], or least loaded server. The answer lies in the extreme dynamicity of cloud environments. Services and load balancers "*must be designed to gracefully withstand traffic surges of hundreds of times their usual loads, as well as DDoS attacks*" [3]. This means that the number of servers and load balancers used to provide a service can quickly change over time. Guaranteeing that packets belonging to existing connections are routed to the correct server despite dynamic reconfigurations requires *per-connection-consistency* (PCC) [32] and has been the focus of many previous works [3, 13, 15, 20, 32, 37, 41]. When only the number of load balancers change, hash-based load balancing mechanisms guarantee PCC as packets reach the correct server even when hitting a different LB [3, 37]. To deal with changes in the numbers of servers, existing LBs either store the "connection-to-server" mapping [13, 20, 32, 41] or let the servers reroute packets that were misrouted [3, 37]. In both cases, a hash function helps mitigate PCC violations, though it cannot completely avoid them (more details in Sect. 2). To summarize, existing load balancers cannot uniformly distribute connections across the servers as they rely on hash functions to mitigate (but not avoid) PCC violations.

This paper presents the design and evaluation of CHEETAH, a load balancer (LB) system with the following properties:

- ***dynamicity***, the number of LBs and servers can increase or decrease depending on the actual load;
- ***per-connection-consistency*** (PCC), packets belonging to the same connection are forwarded to the same server;
- ***uniform load distribution***, by supporting advanced load balancing mechanisms that efficiently utilize the servers;
- ***efficient packet processing***, the LB should have minimal impact on communication latency; and
- ***resilience***, it should be hard for a client to "clog" the LB

and the servers with spurious traffic.

CHEETAH takes a different approach compared to existing LBs. CHEETAH stores information about the connection mappings into the connections themselves. More specifically, when a CHEETAH LB receives the first packet of a connection, it encodes the selected server's identifier into a *cookie* that is permanently added to all the packet headers exchanged within this connection. Unlike previous work, which relies on hash computations to mitigate PCC violations, the design of CHEETAH completely *decouples* the load balancing logic from PCC support. This in turn allows an operator to guarantee PCC regardless of the "connection-to-server" mapping produced by the chosen load balancing logic. The goal of this paper is not the design of a novel load balancing mechanism for uniformly spreading the load but rather the design of CHEETAH as a building block to support PCC for *any* realizable load balancing mechanisms *without* violating PCC. As for resilience, we cannot expose the server identifiers to users as this would open the doors to clogging a targeted server. CHEETAH is designed with resilience in mind, thwarting resource exhaustion and selective targeting of servers. To this end, CHEETAH generates "opaque" cookies that can be processed fast and can only be interpreted by the LB.

We present two different implementations of CHEETAH, a stateless and a stateful version. Our stateless and stateful CHEETAH LBs carefully encode the connection-to-servers mappings into the packet headers so as to guarantee levels of resilience that are no worse (and in some cases even stronger) than existing stateless and stateful LBs, respectively. For instance, our stateful LB increases resilience by utilizing a novel and fast stack-based mechanism that dramatically simplifies the operation of today's cuckoo-hash-based stateful LBs, which suffer from slow insertion times.

In summary, our contributions are:

- We quantify limitations of existing stateless and stateful LBs through large-scale simulations. We show that the quality of the load distribution of existing LBs is 40 times worse than that of an ideal LB. We also show stateless LBs (such as Beamer and Faild) can reduce such imbalances at the price of increasing PCC violations.
- We introduce CHEETAH, an LB that guarantees PCC for any realizable load balancing mechanisms. We present a stateless and a stateful design of CHEETAH, which strike different trade-offs in terms of resilience and performance.
- We implement our stateless and stateful CHEETAH LBs in FastClick [5] and compare their performance with state-of-the-art stateless and stateful LBs, respectively. We also implement both versions of CHEETAH with a weighted round-robin LB on a Tofino-based switch [6].
- In our experiments, we show the potential benefits of CHEETAH with a non-hash-based load balancing mechanism. The number of processor cycles per packet for *both* our stateless and stateful implementation of CHEETAH is comparable to existing stateless implementations and 3.5x less than existing stateful LBs.

2 Background and Motivation

Internet organizations deploy large-scale applications using clusters of servers located within one or more datacenters (DCs). We provide a brief background on DC load balancers, discuss related work, and show limitations of the existing schemes. We do not discuss geo-distributed load balancing across DCs. Further, we distinguish between *stateless* LBs, which do not store any per-connection state, and *stateful* LBs, which store some information about ongoing connections.

Multi-tier load balancing architectures. Datacenter operators assign a *Virtual IP* (VIP) address to each operated service. Each VIP in a DC is associated with a set of servers providing that service. Each server has a *Direct IP* (DIP) address that uniquely identifies the server within the DC.

A LB inside the DC is a device that receives incoming connections for a certain VIP and selects a server to provide the requested service. Each connection is a Layer 4 connection (typically TCP or QUIC). For each VIP, a LB partitions the space of the connection identifiers (*e.g.*, TCP 5-tuples) across all the servers (*i.e.*, DIPs) associated with that VIP. The partitioning function is stored in the LB and is used to retrieve the correct DIP for each incoming packet.

A large-scale DC may have tens of thousands of servers and hundreds of LBs [13, 15, 32]. These LBs are often arranged into different tiers (see Fig. 1). The 1st-tier of LBs are faster and less complex than those in subsequent tiers. For example, a typical DC would use BGP routers using ECMP forwarding at the 1st-tier, followed by Layer 4 LBs, in turn followed by Layer 7 LBs and applications [20]. Similar to previous work on DC load balancing, we consider Layer 7 LBs to be at the same level as the services [13, 20, 41]. Any 1st tier LB receiving a packet directed to a VIP, performs a look up to fetch the *set* of 2nd tier LBs responsible for that VIP. It then forwards the packet towards any of these LBs. The main goal of the 1st-tier is demultiplexing the incoming traffic at the per VIP level towards their dedicated 2nd tier LBs. The 2nd-tier LBs perform two crucial operations: *(i)* guaranteeing (PCC) [32] and *(ii) load balancing* the incoming connections.

2.1 Limits of Stateless Load Balancers

Traditional stateless LBs cannot guarantee PCC. A stateless LB partitions the space of connection identifiers among the set of servers. The partitioning function is stored in the LB and does not depend on the number of active connections. Most stateless LBs, *e.g.*, ECMP [9, 19] & WCMP [53], store this partitioning in the form of an *indirection table*, which maps the output of a hash function modulo the size of the table to a specific server [3, 19, 37, 53].

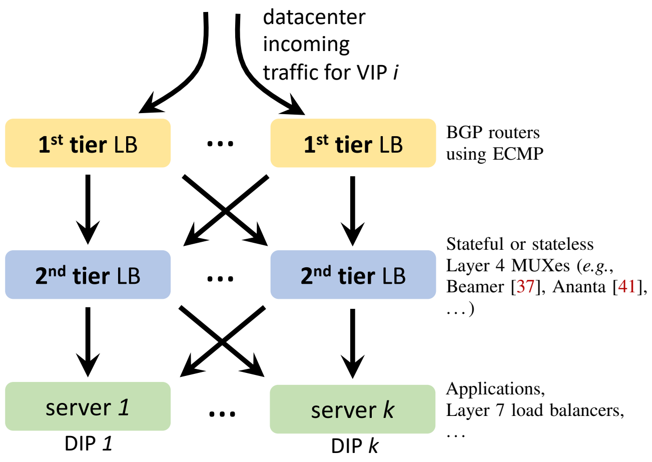


Figure 1: A traditional datacenter load balancing architecture.

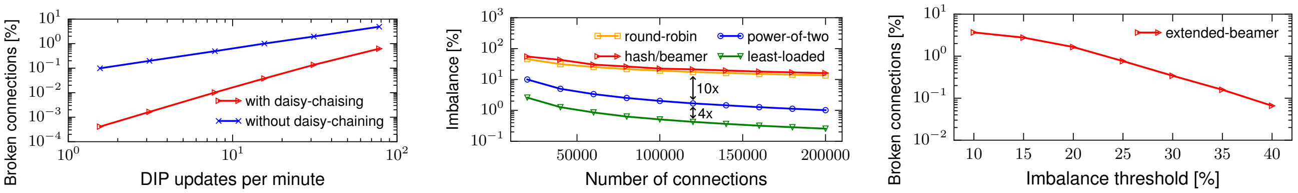
A *uniform hash* scheme maps each server to an equal number of entries in the indirection table. When a LB receives a packet, it extracts the connection identifier from the packet and feeds it as input to a hash function. The output of the hash modulo the size of the indirection table determines the index of the entry in the table where the LB can find to which server the packet should be forwarded. If the number of servers changes, the indirection table must be updated, which may cause some *existing* connections to be rerouted to the new (and wrong) server that is now associated with an entry in the table, *i.e.*, a PCC violation.

Advanced stateless LBs cannot always guarantee PCC. Beamer [37] and Faild [3] introduced *daisy-chaining* to tackle PCC. They encapsulate in the header of the packet the address of a "backup" server to which a packet should be sent when the LB hits the wrong server. This backup server is selected as the last server that was assigned to a given entry in the indirection table before the entry was remapped. PCC violations are prevented as long as *(i)* one does not perform two reconfigurations that change the same entry in the table twice (as only one backup server can be stored in the packet) and *(ii)* one can *simultaneously* reconfigure all the LBs (see [37] for an example).

Fig. 2a shows the percentage of broken connections (*i.e.*, PCC violations) with and without daisy chaining in our large-scale simulations. We used the same parameters, traffic workloads, and cluster reconfiguration events derived from previous work on real-world DC load balancing, *i.e.*, SilkRoad [32]. Namely, we simulated a cluster of 468 servers and we generated a workload using the same traffic distribution of a large web server service. We performed *DIP updates*, *i.e.* removal or additional of servers from the cluster, using different frequency distributions. SilkRoad reports that 95% of their clusters experience between 1.5 and 80 DIP updates/minute and provide distributions for the update time. We define the number of *broken connections* as the number of connections that have been mapped to *at least two* different servers during their starting and ending times. Fig. 2a shows that Beamer and Faild (plotted using the same line) still break almost 1% of the connections at the highest DIP update frequency, which may lead to an unacceptable level of service level agreement (SLA) violations [32].

Hash-based LBs cannot uniformly spread the load. We now investigate the ability of different load balancing mechanisms to uniformly spread the load across the servers for a single VIP. Similarly to the Google Maglev work [13], we define the *imbalance* of a server as the ratio between the number of connections active on that server and the average number of active connections across all servers. We also define the *system imbalance* as the maximum imbalance of any server. The imbalance of a simulation run is the *average* imbalance of the system during the entire duration of the simulation. We discuss different load metrics in Sect. 4. Using the same simulation settings as described above, we compare *(i)* Beamer [37]/Faild [3], which use a uniform hash, *(ii)* Round-Robin [50], which assigns each new connection to the next server in a list, *(iii)* Power-Of-Two [33], which picks the least loaded among two random servers, and *(iv)* Least-Loaded [50], which assigns each new connection to the server with fewest active connections. We note that Round-Robin, Power-Of-Two, and Least-Loaded require storing the connection-to-server mapping, hence they cannot be supported by Beamer/Faild. In this simulation, we do not change the size of the cluster but rather vary only the number of connections that are active at the same time in the cluster between 20K and 200K. We choose this range of active connections to induce the same imbalances (15%-30%) observed for uniform hashes in Google Maglev [13]. Fig. 2b shows the results of our simulations. We note that Beamer-like hash-based LBs outperform consistent hashing by a factor of 2x. Round-Robin outperforms a Beamer-like LB by a factor of 1.2x. When comparing these schemes with Power-Of-Two, we observe a reduction in imbalance by a factor of 10x. Finally, Least-Loaded further reduces the imbalance by an additional factor of 4*x*. These results show that a more uniform distribution of loads can be achieved by storing the mapping between connections and servers, though one still has to support PCC when the LB pool size changes. We note that today's stateful LBs [13,20,32,41] rely on different variations of uniform-hash, thus suffer from imbalances similarly to Beamer.

Beamer can reduce imbalance at the cost of a greater number of PCC violations. We tried to reduce the imbalances in Beamer by monitoring the server load imbalances and modify the entries in the indirection table accordingly. We extended Beamer with a dynamic mechanism that gets as input an imbalance threshold and remove a server from the indirection table whenever its load is above this threshold. The server is



(a) Impact of daisy-chained during DIP updates.

(b) Load imbalance.

(c) Dynamic Beamer imbalance.

Figure 2: Analysis of PCC-violations and load imbalances of state-of-the-art load balancers. To ease visibility, points are connected with straight lines along the x-axis.

re-added to the table when its number of active connections drops below the average. Note that, if an entry in the indirection table changes its server mapping twice, Beamer will break those existing connections that were relying on the initial state of the indirection table. Fig. 2c shows the percentage of broken connections for increasing imbalance thresholds. We set the number of active connections to 70K (corresponding to an *average* 30% imbalance in Fig. 2b). We note that guaranteeing an imbalance of at most 10% would cause 3% of all connections to break. Even with an imbalance threshold of 40% one would still observe 0.1% broken connections because of micro-bursts. Hence, even this extended Beamer cannot guarantee PCC and uniform load balancing at the same time.

2.2 Limits of Stateful Load Balancers

Stateful LBs store the connection-to-server mapping in a so-called `ConnTable` for two main reasons: *(i)* to preserve PCC when the number of servers changes and *(ii)* to enable fine-grain visibility into the flows.

Today's stateful LBs cannot guarantee PCC. Consider Fig. 1 and the case in which we add an additional stateful LB for a certain VIP. The BGP routers, which rely on ECMP, will reroute some connections to a LB than does not have the state for that connection. Thus, this LB does not know to which server the packet should be forwarded unless all LBs use an identical hash-based mechanism (and therefore experience imbalances). Therefore, existing LBs (including Facebook Katran [20], Google Maglev [13], and Microsoft Ananta [41]) rely on hashing mechanisms to mitigate PCC violations. However, this is not enough if the number of servers also changes, then some existing connections will be routed to an LB without state, hence it will hash the connection to the wrong server, thus breaking PCC.

Today's stateful LBs rely on complex and slow data structures. State-of-the-art LBs rely on cuckoo-hash tables [40] to keep per-connection mappings. These data structures guarantee constant time lookups but may require non-constant insertion time [43]. These slow insertions may severely impact the LB's throughput, *e.g.*, a throughput loss by 2x has been observed on OpenFlow switches when performing ~ 60 updates/second [34].

2.3 Service Resilience and Load Balancers

Load balancers are an indispensable component against clogging Distributed Denial of Service (DDoS) attacks, *e.g.*, bandwidth depletion at the server and memory exhaustion at the LB. Dealing with such attacks is a multi-faceted problem involving multiple entities of the network infrastructure [30], *e.g.*, firewalls, network intrusion detection, application gateways. This paper does not focus on how the LB fits into this picture but rather studies the resilience of the LB itself and the resilience its design provides to the service operation.

LBs shield servers from targeted bandwidth depletion attacks. An LB system should guarantee that the system absorb sudden bursts due to DDoS attacks with minimal impact on a service's operation. Today's LB mechanisms rely on hash-based load balancing mechanisms to provide a first pro-active level of defense, which consists in spreading connections across all servers. As long as an attacker does not reverse engineer the hash function, multiple malicious connections will be spread over the servers. A system should not allow clients to target specific servers with spurious traffic.

Stateful LBs support per-connection view at lower resilience. Stateful LBs provide fine-grained visibility into the active connections, providing resilience to the service operation, *e.g.*, by selectively rerouting DDoS flows. At the same time, stateful LBs are a trivial target of resource depletion clogging DDoS attacks: incoming spurious connections add to the connection table rapidly exhaust the limited LB memory (*e.g.*, [37]) or grow the connection table aggressively, rapidly degrading performance even with ample memory [34]. Stateless LBs can inherently withstand clogging DDoS, sustaining much higher throughput, but can only offer per-server statistics visibility to the service operation.

Having analyzed the above limitations of existing load balancers, we conclude this section by asking the following question: *"Is it possible to design a DC load balancing system that guarantees PCC, supports any realizable load balancing mechanism, and achieves similar levels of resiliency of today's state-of-the-art LBs?"*

3 The Cheetah Load Balancer

In this section, we present Cheetah, a load balancing system that supports arbitrary load balancing mechanisms and guarantees PCC without sacrificing performance. Cheetah solves many of the today's load balancing problems by encoding information about the connection into a *cookie* that is added to all the packets of a connection.[1] Cheetah sets the cookie according to any chosen and realizable load balancing mechanism and relies on that cookie to *(i)* guarantee future packets belonging to the same connection are forwarded to the same server and *(ii)* speed up the forwarding process in a stateful LB, which in turn increases the resilience of the LB. Understanding what information should be encoded into the cookie, how to encode it, and how to use this information inside a stateless or stateful LB is the goal of this section. We start our discussion by introducing the stateless Cheetah LB, which guarantees PCC and preserves the same resilience and packet processing performance of existing stateless LBs. We then introduce the stateful Cheetah LB, which improves the packet processing performance of today's stateful LBs, and present an LB architecture that strikes different tradeoffs in terms of performance and resilience. We stress the fact that Cheetah does not propose a novel LB mechanism but is a building block for supporting arbitrary LB mechanisms without breaking PCC (we show the currently implemented LB mechanisms in Sect. 4).

A naïve approach. We first discuss a straightforward approach to guarantee PCC that would not work in practice because of its poor resiliency. It entails storing the identifier of a server (*i.e.*, the DIP) in the cookie of a connection. In this way, an LB can easily preserve PCC by extracting the cookie from each subsequent incoming packet. We note that such naïve approaches are reminiscent of several previous proposals on multi path transport protocols [10, 39], where the identifiers of the servers are explicitly communicated to the clients when establishing multiple subflows within a connection. There is at least one critical resiliency issue with this approach. Some clients can wait to establish many connections to the same server and then suddenly increase their load. This is highly undesired as it leads to cascade-effect imbalances and service disruptions [47].

3.1 Stateless Cheetah LB

The stateless Cheetah LB: encoding an opaque offset into the cookie. We now discuss how we overcome the above issues in Cheetah. We aim to achieve the same resiliency levels[2] of today's production-ready stateless LBs (*e.g.*, Faild [3]/Beamer [37, 47]) while supporting arbitrary load balancing mechanisms and guaranteeing PCC. We assume a single tier LB architecture and defer the discussion of multi-tier architectures to later in this section.

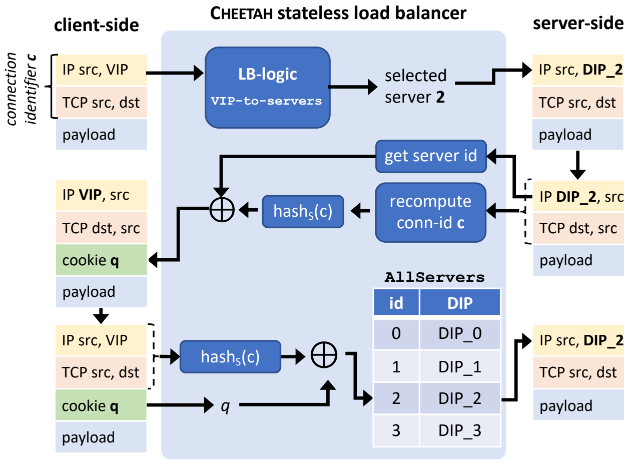


Figure 3: Cheetah stateless LB operations.

The Cheetah stateless LB keeps two different types of tables (see Fig. 3): an `AllServers` table that maps a server identifier to the DIP of the server and a `VIPToServers` table that maps each VIP to the set of servers running that VIP. The `AllServers` table is mostly static as it contains an entry for each server in the DC network. Only when servers are deployed in/removed from the DC is the `AllServers` table updated. The `VIPToServers` table is modified when the number of servers running a certain service increases/ decreases, a more common operation to deal with changes in the VIP current demands.

When the LB receives the first packet of a connection (top part of Fig. 3), it extracts the set of servers running the service (*i.e.*, with a given VIP) from the `VIPToServers` table, selects one of the servers according to any pre-configured load balancing mechanism, and forwards the packet.[3] For every packet received from a server (middle part of Fig. 3), the LB encodes an "opaque" identifier of the server mapping into the cookie for this connection. To do so, Cheetah computes the hash of the connection identifier with a salt S (unknown to the clients), XORs it with the identifier of the server, and adds the output of the XOR to the packet header as the cookie. The salt S is the same for all connections. When the LB receives any subsequent packet belonging to this connection (bottom part of Fig. 3), it extracts the cookie from the packet header, computes the hash of the connection identifier with the salt S, XORs the output of the hash with the cookie, and uses the output of the XOR as the identifier of the server. The LB then looks up the DIP of the server in the `AllServers` table.

Stateless Cheetah guarantees PCC. Cheetah relies on two main design ideas to avoid breaking connections:

[1] We discuss legacy-compatibility issues in Sect. 4

[2] See Sect 2.3 for details of stateless/stateful load balancing resiliency.

[3] How to implement different LB mechanisms in programmable hardware and software LBs is shown in Sect. 4.

(i) moving the state needed to preserve the mapping between a connection and its server into the packet header of the connection and *(ii)* using the more dynamic `VIPToServers` table only for the 1st packet of a connection. Subsequently, the static `AllServers` table is used to forward packets belonging to any existing connection. This trivially guarantees PCC. We defer discussion of multi-path transport protocols to Sect. 6.

Compared to existing stateless LBs Stateless CHEETAH achieves similar resiliency. Binding the cookie with the hash of the connection identifier brings one main advantage compared to the earlier naïve scheme, as an attacker must first reverse engineer the hash function of the LB in order to launch an attack targeting a specific server. This makes CHEETAH as resilient as other production-ready stateless LBs. We note that CHEETAH is orthogonal to DDoS mitigation defence mechanisms, especially when deployed in reactive mode. We further discuss CHEETAH resilience, including support for multi path transport protocols, in Sect. 6.

Stateless CHEETAH supports arbitrary load balancing mechanisms. All the reviewed state-of-the-art LBs (even stateful ones) are restricted to uniform hashing when it comes to load balancing mechanisms — as any other mechanism would break an unacceptable number of connections when the number of servers/LBs changes. In contrast, whenever a new connection arrives at a stateless CHEETAH LB, CHEETAH selects a server among those returned from a lookup in the `VIPToServers` table. The selected server may depend upon the specific load balancing mechanism configured by the service's operator. We note that the selection of the server may or may not be implementable in the data-plane. The CHEETAH LB guarantees that once the mapping connection-to-server has been established by the LB logic (not necessarily at the data-plane speed), all the subsequent packets belonging that that connection will be routed to the selected server. Since the binding of the connection to the server is stored in the packet header, CHEETAH can support LB mechanisms that go well beyond uniform hashing. For instance, an operator may decide to rely on "power of two choices" [33], which is known to reduce the load imbalance by a logarithmic factor. Another service operator may prefer a weighted round-robin load balancing mechanism that uses some periodically reported metrics (*e.g.*, CPU utilization) to spread the load uniformly among all the servers.

Lower bounds on the size of the cookie. In CHEETAH, the size of the cookie has to be at least $\log_2 k$ bits, where k is the maximum number of servers stored in the `AllServers` table. Therefore, the size of the cookie grows logarithmically in the size of the number of servers. One question is whether PCC can be guaranteed using a cookie whose size is smaller than $\log_2(k)$ and the memory size of the LB is constant. We defer proof of the following theorem to App. A.

Theorem 1. Given an arbitrarily large number of connections, any load balancer using $O(1)$ memory requires cookies of size $\Omega(\log(k))$ to guarantee PCC under any possible change in the number of active servers, where k is the overall number of servers in the DC that can be assigned to the service with a given VIP.

In App. A, we generalize the above theorem to show a certain class of advanced load balancing mechanisms, including round-robin and least-loaded, requires cookies with a size of at least $\log_2(k)$ bits even in the absence of changes in the set of active servers.

While the above results close the doors to any sublogarithmic overhead in the packet header; in practice, operators may decide to trade some PCC violations and load imbalances for a smaller sized cookie. We refer the reader to App. B for a discussion about how to implement CHEETAH with limited size cookies.

3.2 Stateful CHEETAH LB

We also designed a stateful version of CHEETAH to support a finer level of visibility into the flows than that offered by stateless LBs. A stateful LB can keep track of the behaviour of each individual connection and support complex network functions, such as rate limiters, NATs, detection of heavy-hitters, and rerouting to dedicated scrubbing devices (as in the case of Microsoft Ananta [41] and CloudFlare [30]). In contrast to existing LBs, our stateful LB guarantees PCC (inherited from the stateless design) and uses a more performant `ConnTable` that is amenable to fast data plane implementations. In the following text we say that PCC is guaranteed if a packet is routed to the correct server as long as an LB having state for its connection exists.

The stateful CHEETAH load balancer: encoding table indices in the packet header. As discussed in Sect. 2, today's stateful LBs rely on advanced hash tables, *e.g.*, cuckoo-hashing [40], to store per-connection state at the LB [32]. Such data structures offer constant-time data-plane lookups but insertion/modification of any entry in the table requires intervention of the slower control plane or complex & workload-dependant data structures (*e.g.*, Bloom filters [32], Stash-based data structures [43]), which are both complex and hard to tune for a specific workload.

We make a simple yet powerful observation about stateful tables that any insertion, modification, or deletion of an entry in a table can be greatly simplified if a packet carries information about the index of the entry in the table where its connection is stored. Since datacenters may have tens of billions of active connections, we need to devise a stateful approach where the size of the cookie is explicitly given as input. In a stateful CHEETAH LB (see Fig. 4), we store a set of m `ConnTable` tables that keep per-connection statistics and DIP mappings. We also use an equal number of `ConnStack` stacks of indices, each storing the unused entries in its corresponding `ConnTable`.

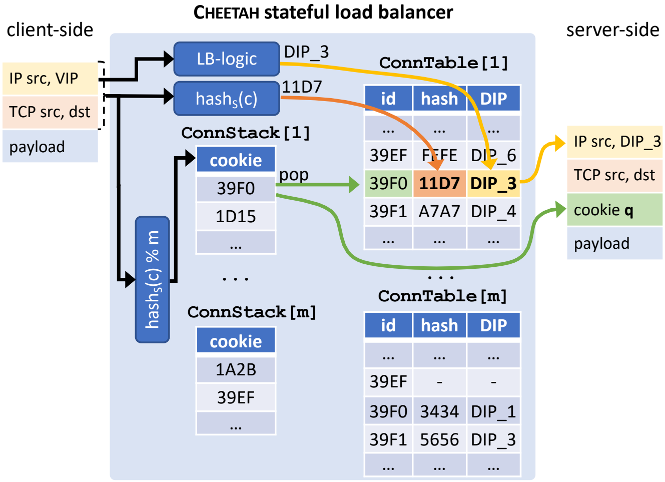


Figure 4: CHEETAH stateful LB operations for the 1st packet of a connection. We do not show the stateless cookie for identifying the stateful LB. The VIP-to-servers is included within the LB-logic and not shown. The server performs Direct Server Return (DSR) so the response packet does not traverse the load balancers. Subsequent packets from the client only access their index in the correspoding ConnTable.

For the sake of simplicity, we first assume there is only one LB and one ConnTable with its associated ConnStack, *i.e.*, $m = 1$. Whenever a new connection state needs to be installed, CHEETAH pops an index from ConnStack and incorporates it as part of the cookie in the packet's header. It also stores the selected server and the hash of the connection identifier with a salt S into the corresponding table entry. This hash value allows the LB to filter out malicious attempts to interfere with legitimate traffic flows, similarly to SilkRoad [32]. Whenever a packet belonging to an existing connection arrives at the LB, CHEETAH extracts the index from the cookie and uses it to quickly perform a lookup only in the ConnTable. Note that insertion, modification, and deletion of connections can be performed in constant time entirely in the data plane. We explain details of the implementation in Sect. 4.

The number of connections that we can store within a single ConnTable is equal to 2^r, where r is the size of the cookie. In practice, the size of the cookie may limit the number of connections that can be stored in the LB. We therefore present a hybrid approach that uses a hash function to partition the space of the connection identifiers into m partitions. As for any stateful table, m should be chosen high enough so the total number of entries $m * 2^r$ is suitable. The same cookie can be re-used among connections belonging to distinct partitions.

A hybrid datacenter architecture. Stateful LBs are typically not deployed at the edge of the datacenter for two reasons: they are more complex and slower compared to stateless LBs. As such, they are a weak point that could compromise the entire LB availability. Therefore, we propose a 2-tier DC architecture where the first tier consists of stateless CHEETAH LBs and the second tier consists of stateful CHEETAH LBs. The stateless LB uses the first bytes of the cookie to encode the identifier of a stateful load balancer, thus guaranteeing a connection always reaches the same LB regardless of the LB pool size. The stateful load balancer uses the last bytes of the cookie to encode per-connection information as described above.

4 Implementation

The simplicity of our design makes CHEETAH amenable to highly efficient implementations in the data-plane. We implemented stateful and stateless CHEETAH LBs on FastClick [5], a faster version of the Click Modular Router [26] that supports DPDK and multi-processing. Previous stateless systems, such as Beamer [37], have also relied on FastClick for their software-based implementation. We also implemented stateless and stateful versions of the CHEETAH LB with a weighted round-robin LB on a Tofino-based switch using P4 [6].[4] We can only make a general P4 implementation available due to Tofino-related NDAs. Both implementations are available at [4]. We first discuss the critical question of where to actually store the cookie in today's protocols and then describe the FastClick and P4-Tofino implementations.

Preserving legacy-compatibility. Our goal is to limit the amount of modifications needed to deploy CHEETAH on existing devices. Ideally, we would like to use a dedicated TCP option for storing the CHEETAH cookie into the packet header of all packets in a connection. However, this would require modifications to the clients, which would be infeasible in practice. We therefore identified three possible ways to implement cookies within existing transport protocols *without* requiring any modifications to the clients' machines: *(i)* incorporate the cookie into the connection-id of QUIC connections, *(ii)* encode the cookie into the least significant bits of IPv6 addresses and use IPv6 mobility support to rebind the host's address (the LB acts as a home agent), and *(iii)* embed the cookie into part of the bits of the TCP timestamp options. In this paper, we implemented a proof-of-concept CHEETAH using the TCP timestamp option as explained in App. C.[5] . We note that similar encodings of information into the TCP timestamp have been proposed in the past but require modifications to the servers [39]. The stateless CHEETAH LB can transparently translates the server timestamps with the encoded timestamps without interfering

[4] Detailed performance benchmarking of CHEETAH on the Tofino switch is subject to an NDA. The Tofino implementations follow the description of the mechanisms presented in Sect. 3, use minimal resources, and incur neither significant performance overheads nor require packet recirculation.

[5] We verified in App. C that the latest Android, iOS, Ubuntu, and MacOS operating systems support TCP timestamp options but not Windows.

with TCP timestamp related mechanisms (*i.e.*, RTT estimation and protections against wrapped sequences [8]). Therefore, no modifications are required to the servers for stateless mode unless the datacenter operator wants to guarantee Direct Server Return (DSR), *i.e.*, packets from the servers to the client do not traverse any load balancer. In that case, the server must encode the cookie into the timestamp itself. The cookie must also be sent back by the server for stateful mode, as the load balancer would not be able to find the stack index for returning traffic. Server modifications are described in C.2. We leave the implementation of CHEETAH on QUIC and IPv6 as future work. We note that a QUIC implementation would be easier and more performant since parsing TCP options is an expensive operation in both software and hardware LBs.

4.1 FastClick implementation

The FastClick implementation is a fully-fledged implementation of CHEETAH that supports L2 & L3 load balancing and multiple load balancing mechanisms (*e.g.*, round-robin, power-of-2 choices, least-loaded server). The LB supports different load metrics including number of active connection and CPU utilization. The LB decodes cookies for both stateless and stateful modes using the TCP timestamp as described above, and can optionally fix the timestamp in-place if the server is not modified to do it.

Parsing TCP options. Each TS option has a 1-byte identifier, 1-byte length, and then the content value. Options may appear in any order. This makes extracting a specific option a non-trivial operation [10]. We focus on extracting the timestamp option TS_{ecr} from a packet. To accelerate this parsing operation, we performed a statistical study over 798M packets headers from traffic captured on our campus.

Table 1 shows the most common patterns observed across the entire trace for packets containing the timestamp option. The Linux Kernel already implements a similar fast parsing technique for non-SYN(/ACK) packets. We first consider non-SYN packets (*i.e.*, “Other packets” in the table). Our study shows that 99.95% of the packets have the following pattern: NOP (1B) + NOP (1B) + TimeStamp (10B) possibly followed by other fields. When a packet arrives, we can easily determine whether it matches this pattern by performing a simple 32-bit comparison and checking that the first two bytes are NOP identiers and the third one is the Timestamp id. We process the remaining 0.05% of the traffic in the slow path. We now look at SYN packets. Consider the first row in the table, *i.e.*, MSS (4B) + SAckOK (2B) + TimeStamp (10B) + SAck + EOL. To verify if a packet matches this pattern, we perform a 64-bit wildcard comparison and check that the first byte is the MSS id, the fifth byte is the SAckOK id, and the seventh byte is the TimeStamp id. We can apply similar techniques for the remaining patterns matchable with 64 bits. Some types of hosts generate packets whose patterns are wider than 64 bits, which is the limit of our x86_64 machine. We then rely on one SSE 128bit integer wildcard comparison to verify such patterns. The remaining 2.24% of patterns are handled through a standard hop-by-hop parsing following the TCP options Type-Length-Value chain. Finally, we note that we can completely avoid the more complex parsing operations for SYNs and SYN/ACKs if servers use TCP SYN cookies [12] (see App. C for more details).

Table 1: TCP Options pattern

SYN packets	
MSS SAckOK Timestamp [NOP WScale]	49.86%
MSS NOP WScale NOP NOP Timestamp [SAckOK EOL]	44.49%
MSS NOP WScale SAckOK Timestamp	4.53%
Slow path	1,12%
SYN-ACK packets	
MSS SAckOK Timestamp [NOP WScale]	76.85%
MSS NOP WScale SAckOK Timestamp	18.79%
MSS NOP NOP Timestamp [SAckOK EOL]	1.69%
MSS NOP WScale NOP NOP Timestamp [SAckOK EOL]	1.55%
Slow path	1,12%
Other packets	
NOP NOP Timestamp	98.46%
NOP NOP Timestamp [NOP NOP SAck]	1.49%
Slow path	0,05%

Load balancing mechanisms. CHEETAH supports any realizable LB mechanisms while guaranteeing PCC. We implemented several load balancing mechanisms that will be evaluated using multiple workloads in Sect. 5.2. Among the load-aware LB mechanisms, we distinguish between metrics that can be tracked with or without coordination. Without any coordination, the LB can keep track of the number of packets/bytes sent per server and an estimate of the number of open connections based on a simple SYN/FIN counting mechanism.[6] For LB approaches that require coordination with the servers, our implementation supports load distribution based on the CPU utilization of the servers. Note that using a least-loaded server for coordination-based approaches is a bad idea as a single server will receive all the incoming connections until its load metric increases and is reported to the LB, ultimately leading to instabilities in the system. Therefore, we decided to implement the following two load-aware balancing mechanisms, which we introduced in Sect. 2: *(i)* power-of-2 choices and *(ii)* a weighted round robin (WRR). For WRR, we devised a system where the weights of the servers change according to their relative (CPU) loads. We increase the weights for servers that are underutilized depending on the difference between their load and the average server load. More formally, the number of buckets N_i assigned to server i is computed as $N_i = round(10\frac{L_{avg}}{(1-\alpha)*L_i+\alpha*L_{avg}})$ where L_i is the load of a server, and α is a factor that tunes the speed of the convergence, which we set to 0.5. A perfectly balanced system would give $N = 10$ buckets to each server. An underutilized server gets more than

[6] We envision an ad-hoc mechanism to signal closed connection between the LB and the server would make the estimate reliable in the future.

N buckets (in practice limited to $3N$) while an overloaded server gets less than N buckets (lower bounded by 2).

4.2 P4-Tofino prototype

The stateless Cheetah LB follows exactly the description from Sect. 3.1. We store the `all-servers` and the `VIP-to-servers` tables using exact-match tables. We rely on registers, which provide per-packet transactional memories, to store a counter that implements the weighted-round-robin LB. We note that implementing other types of LB mechanisms such as least-loaded in the data-plane is non trivial in P4 since one would need to extract a minimum from an array in $O(1)$. This operation will likely requires to process the packet on the CPU of the switch. The insertion/deletion of the cookie on any subsequent non-SYN packet can be performed in the data-plane. The stateful Cheetah LB adheres to the description in Sect. 3.2. We use P4 registers to enable the insertion of connections into the `ConnTable` at the speed of the data-plane. We store the elements of the `ConnStack` stack in an array of registers, the `ConnTable` into an array of registers, and the pointer to the head of the stack in another register.

5 Evaluation

The Cheetah LB design allows datacenter operators to unleash the power of arbitrary load balancing mechanisms while guaranteeing PCC, *i.e.*, the ability to grow/shrink the LB and DIP pools without disrupting existing connections. In this section, we perform a set of experiments to assess the performance achievable through our stateless and stateful LBs. We focus only on evaluating the performance of the FastClick implementation.[7] All experiments scripts, including documentation for full reproducibility are available at [4].

We pose three main questions in this evaluation:

- *"How does the **cost of packet processing** in* Cheetah *compare with existing LBs?"* (Sect. 5.1)
- *"Can we reduce **load imbalances** by implementing more advanced LB mechanisms in* Cheetah*?"* (Sect. 5.2)
- *"How does the **PCC support** in* Cheetah *compare with existing stateless LBs?"* (Sect. 5.3)

Experimental setting. The LB runs on a dual-socket, 18-core Intel®Xeon®Gold 6140 CPU @ 2.30GHz, though only 8 cores are used from the socket attached to the NIC. Our testbed is wired with 100G Mellanox Connect-X 5 NICs [48] connected to a 32x100G NoviFlow WB5000 switch [36]. All CPUs are fixed at their nominal frequency.

Workload generation. To generate load, we use 4 machines with a single 8-core Intel®Xeon®Gold 5217 CPU @ 3.00GHz with hyper-threading enabled using an enhanced version of WRK [17] to generate load towards the LB. We also use four machines to run up to 64 NGINX web servers (one per hyper-thread), isolated using Linux network namespaces. Each NGINX server has a dedicated virtual NIC using SRIOV, allowing packets to be switched in hardware and directly received on the correct CPU core. We generate requests from the clients using uniform and bimodal distributions, as well as the large web server service distributions already used in the simulations of Sect. 2.

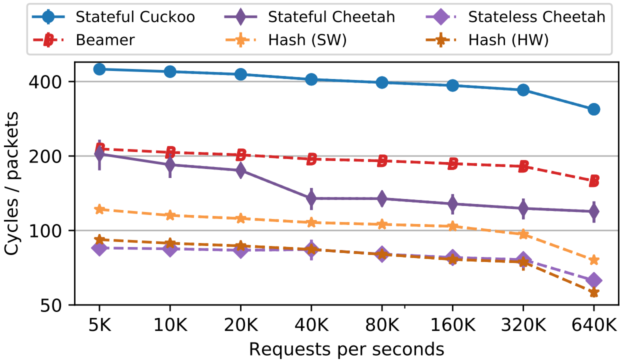


Figure 5: CPU cycles/packets for various methods. Cheetah achieves the same load balancing performance as stateful LBs with 5x fewer cycles and only a minor penalty over hashing.

Metrics. We evaluate the imbalance among servers using both the variance of the server loads and the 99th percentile flow completion times (FCTs), where the latter one is key for latency-sensitive user applications. We measure the LB packet processing time in CPU cycles per second. Each point is the average of 10 runs of 15 seconds unless specified otherwise.

5.1 Packet Processing Analysis

We first investigate the cost in terms of packet processing time for using stateless Cheetah. We compare it against stateful Cheetah, a stateful LB based on per-core DPDK cuckoo-hash tables, and two hashing mechanisms, one using the hash computed in hardware by the NIC for RSS [21], and one computed in software with DPDK [29]. We also compare with a streamlined version of Beamer [38], without support for bucket synchronization, UDP, and MPTCP, thus representing a lower-bound on the Beamer packet processing cost.

Stateless Cheetah incurs minimal packet processing costs. Fig. 5 shows the number of CPU cycles consumed by different LBs divided by the number of forwarded packets for increasing number of requests per second. We tune the request generation for a file of 8KB so that none of the clients or servers were overloaded. The main result from this experiment is that stateless Cheetah consumes almost the same number of CPU cycles per packet as the most optimized, hardware assisted hash-based mechanism and significantly fewer cycles than stateful approaches. Beamer consumes more cycles than both Cheetah LBs, still without bringing PCC

[7] We argue that our Tofino implementation would perform similarly in terms of ability to uniformly distribute the load.

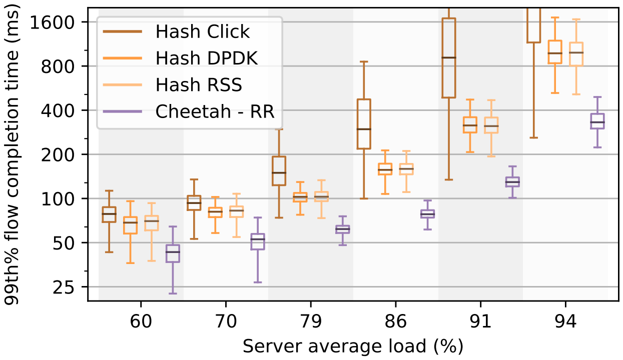


Figure 6: 99th-perc. FCT for the increasing average server load. CHEETAH achieves 2x−3x lower FCT than Hash RSS.

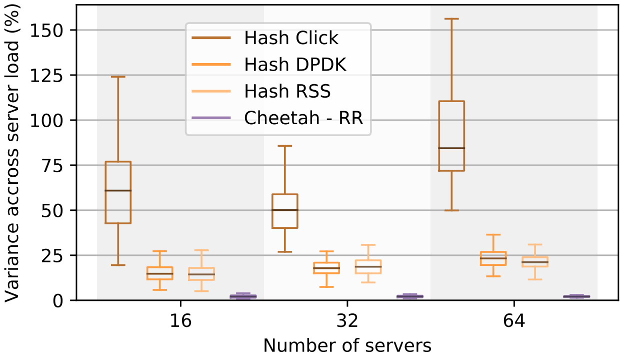


Figure 7: Variance among servers' load of various methods for an increasing number of servers. The average requests/s is 100 per server. CHEETAH, though stateless, allows a near-perfect load spreading.

guarantee (see Sect. 5.3). This is mainly due to the operation of encapsulating the backup server into the packet header and the more compute-intensive operations needed by Beamer to lookup into a bigger "stable hashing" table. Finally, we note that, with the web service requests size distribution, each methods only need 4 CPU cores to saturate the 100Gbps link.

Stateful CHEETAH outperforms cuckoo-hash based LBs. We also note in Fig. 5 the improvements in packet processing time of stateful CHEETAH (which uses a stack-based `ConnTable` table) compared to the more expensive stateful LBs using a cuckoo-hash table. Stateful CHEETAH achieves performance close to a stateless LB and a factor of $2-3$x better that cuckoo-hash based LBs.

Dissecting stateless CHEETAH performance. The key insight into the extreme performance of CHEETAH is that the operation of obfuscating the cookie only adds less than a 4-cycle hit. We in fact rely on the network interface card hardware to produce a symmetric hash (*i.e.*, using RSS). We expect the advent of SmartNICs as well as QUIC and IPv6 implementations, which have easier-to-parse headers, to perform even better. We note that our stateless CHEETAH implementation uses server-side TCP timestamp correction (see Sect. 4), which only imposes a 0.2% performance hit over the server processing time. If we were to use LB-side timestamp correction, we observe that the stateless CHEETAH modifies the timestamp MSB on the LB in just 30 cycles per packet performance hit. To summarize, stateless CHEETAH brings the same benefits as stateful LBs (in terms of load balancing capabilities) in addition to PCC guarantees at basically the same cost (and resilience) of stateless LBs.

5.2 Load Imbalance Analysis

We now assess the benefits of running CHEETAH using a non-hash-based load balancing mechanism and compare it to different uniform hash functions (similarly to those implemented in Microsoft Ananta [41], Google Maglex [13], Beamer [37], and Faild [3]). We stress that we do not propose novel load balancing mechanisms but rather showcase the potential benefits of a load balancer design that supports any realizable load balancing mechanisms. We only evaluate stateless CHEETAH as the load imbalance does not depend on the stored state (and would result in similar performance).

In this experiment, each server performs a constant amount of CPU-intensive work to dispatch a 8KB file. The generator makes between 100 and 200 requests per server per second on average depending on our targeted system load. Given this workload and service type, we expect an operator to choose a uniform round-robin LB mechanism to distribute the load.

CHEETAH significantly improves flow completion time. Fig. 6 compares CHEETAH with round-robin and hash-based LB mechanisms with 64 servers. We consider three hash functions: Click [26], DPDK [29], and the hardware hash from RSS. We stress the fact that these hash-based functions represent the quality of load balancing achievable by existing stateless (*e.g.*, Beamer [37]) and stateful LBs (*e.g.*, Ananta [41]) LBs. We measure the 99th percentile flow completion time (FCT) tail latency for the increasing average server load. We note that CHEETAH reduces the 99th percentile FCT by a factor of $2-3$x compared to the best performing hash-based mechanism, *i.e.*, Hash RSS.

CHEETAH spreads the load uniformly. To understand why CHEETAH achieves better FCTs, we measure the variance of the servers' load over the experiment for an average server load of 60% and 16, 32, and 64 servers. Fig. 7 shows that (as expected) the variance of RR is considerably smaller than hash-based methods. This is because the load balancer iteratively spreads the incoming requests over the servers instead randomly spreading them. In this specific scenario, CHEETAH allows operators to leverage RR, which would otherwise be impossible with today's load balancers. Fig. 7 also shows that the quality of the hash function is important

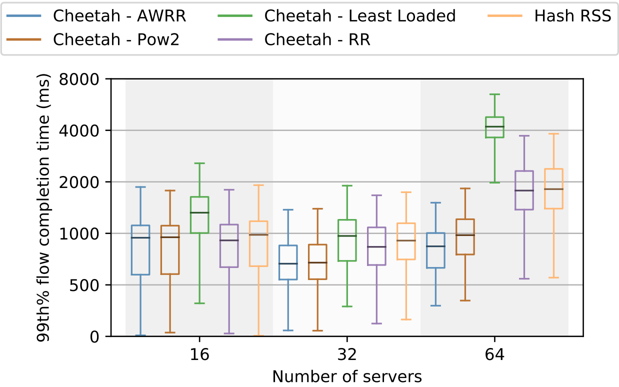


Figure 8: Evaluation of multiple load balancing methods for a bimodal workload. Both AWRR and Pow2 outperform Hash RSS by a factor of 2.2 and 1.9 respectively with 64 servers.

as the default function provided in Click does not perform well. In contrast, the CRC hash function used by DPDK is comparable to the Toeplitz based function used in RSS [28]. Moreover, the RSS function has the advantage of being performed in hardware.

CHEETAH improves FCT even with non-uniform workloads. Fig. 8 shows the tail FCT for a bimodal workload, where 10% of requests take 500ms to be ready for dispatching and the remaining ones take a few hundred microseconds. In this scenario, some servers will be loaded in an unpredictable way thus creating a skew that requires direct feedback from the servers to solve. We can immediately see that RR with 64 servers leads to very high FCTs. We evaluate three ways to distribute the incoming requests according to the current load (see Sect. 4.1): automatic weighted round robin (AWRR), power of two choices (Pow2), and the least loaded server. Each server piggybacks its load using a monitoring Python agent on the server that reports its load through an HTTP channel to the LB at a frequency of 100Hz, though experimental results showed similar performance at 10Hz. Least loaded performs poorly since it sends all the incoming requests to the same server for 10ms, overloading a single server. Pow2 and AWRR spread the load more uniformly as the LB penalizes those servers that are more overloaded. Consequently, both methods reduce the FCT by a factor of two compared to Hash RSS with 64 servers. These experiments demonstrate the potential of deploying advanced load balancing mechanisms to spread the service load.

5.3 PCC Violations Analysis

We close our evaluation by demonstrating the key feature of the stateless CHEETAH LB, *i.e.*, its ability to avoid breaking connections while changing the server and/or LB pool sizes. We compare CHEETAH against Hash RSS, consistent hashing, and Beamer. We start our experiment with a cluster consisting of 24 servers. We tune a python generator [4] to create 1500 requests/s, increasing following a sinusoidal load to 2500 requests/s and descending back to 1500 over the 40 seconds of the experiment. The size and duration of the requests are served using the web server distribution. We iteratively add 7 servers to the pool as the load increases. We then drain 8 servers when the rate goes down. Fig. 9 shows the percentage of broken requests over completed requests every second over time. Some connections gets accounted as broken dozen of seconds later as clients continue sending retransmission before raising an error. Compared to Beamer, Cheetah not only achieves better load balancing with AWRR (Sect. 5.2), but it also does not break any connection.

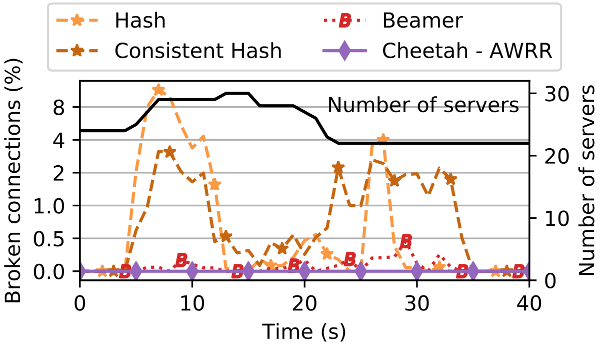


Figure 9: Percentage of broken requests while scaling the number of servers. Cheetah guarantees PCC whereas hashing breaks up to 11% of the connections, consistent hashing 3% and Beamer up to 0.5%.

6 Frequently Asked Questions

Does CHEETAH preserve service resilience compared to existing LBs? Yes. We first discuss whether a client can clog a server. A client generating huge amounts of traffic using the same connection identifier can be detected and filtered out using heavy-hitter detectors [41]. This holds for any stateless LBs, *e.g.*, Beamer [37]. A more clever attack entails reverse engineering the salted hash function and deriving a large number of connection identifiers that the LB routes to the same (specific) server, possibly with spoofed IP addresses. To do so, an attacker needs to build the $(conn.id, cookie) \mapsto server$ mapping. This requires performing complex measurements to verify whether two connection IDs map to the same server. Given that CHEETAH uses the same hash function of any existing LB (which is not cryptographic due to their complexity [3]), reverse engineering this mapping will be as hard as reverse engineering the hash of the existing LBs. As for the resilience to resource depletion, we note from Fig. 5 that stateless (stateful) CHEETAH has similar (better) packet processing times of today's stateless (stateful) LBs. Thus, we argue that CHEETAH achieves the same levels of resilience of today's existing LB systems.

Does CHEETAH make it easy to infer the number of servers? Not necessarily. A 16-bit cookie permits at least an order of magnitude more servers than the number of servers used to operate the largest services [32]. If this is still a concern, one can hide the number of servers by reducing the size of the cookie and partitioning the connection identifier space similarly to our stateful design of CHEETAH.

Does CHEETAH support multipath transport protocols? Yes. In multipath protocols, different sub-connection identifiers must be routed to the same DIP. Previous approaches exposed the server's id to the client [10, 39]; however, this decreases the resilience of the system decreases. CHEETAH can use a different permutation of `AllServers` for each additional *i*'th sub-connection. Clients inform the server of the new sub-connection identifier to be added to an existing connection. The server replies with the cookie to be used using the *i*'th `AllServers` table. This keeps the resilience of the system unchanged compared to the single path case.

7 Related Work

There exists a rich body of literature on datacenter LBs [2, 7, 11, 13, 16, 18–20, 22–25, 32, 37, 41, 52, 53]. We do not discuss network-level DC load balancers [2, 7, 16, 18, 22, 24, 25, 52], whose goal is to load balance the traffic within the DC network and do not deal with per-connection-consistency problems.

Stateless LBs. Existing stateless LBs rely on hash functions and/or "daisy chaining" techniques to mitigate PCC violations (2), *e.g.*, ECMP [19], WCMP [53], consistent hashing [23], Beamer [37], and Faild [3]. The main limitation of such schemes is the suboptimal balancing of the server loads achieved by the hash function, which is known to grow exponentially in the number of servers [49]. Shell [42] proposed a similar use of the timestamp option as a reference to an history of indirection tables, which comes at both the expense of memory and low-frequency load rebalancing.

QUIC-LB [11] is a high-level design proposal at the IETF for a stateless LB that leverages the `connection-id` of the QUIC protocol for routing purposes. While sharing some similarities to our approach, QUIC-LB *(i)* does not present a design of a stateful LB that would solve cuckoo-hash insertion time issues, *(ii)* does not evaluate the performance obtainable on the latest generation of general-purpose machines, *(iii)* relies on the modulo operation with an odd number to hide the server from the client, an operation that is not supported in P4, and *(iv)* does not discuss multi path protocols. We note that our cookie can be implemented as the 160-bit `connection-id` in QUIC, which is also easier to parse than the TCP timestamp option. Encoding the connection-to-server mapping has recently been briefly discussed in an editorial note without discussing LB resilience, stateful LBs, or implementing and evaluating such a solution [31].

Several stateless load balancers that support multi path transport protocols have been proposed in the past. Such load balancers guarantee all the subflows of a connection are routed to the same server by explicitly communicating an identifier of the server to the client [10, 39]. These approaches may be exploited by malicious users to cause targeted imbalances in the system, which is prevented in CHEETAH thanks to using distinct hashes for the subflows (see Sect. 6).

Stateful LBs. Existing stateful LBs store the connection-to-server mapping in a cuckoo-hash table [13, 15, 20, 32, 41] (see Sect. 2). These LBs still rely on hash-based LB mechanisms — as these lead to fewer PCC violations when changing the number of LBs. In contrast, CHEETAH decouples PCC support from the LB logic, thus allowing operators to choose any realizable LB mechanism. Moreover, cuckoo-hash tables suffer from slow (non-constant) insertion time. FlowBlaze [43] and SilkRoad [32] tackled this problem using a stash-based and bloom-filter-based implementations, respectively. Yet, both solutions cannot guarantee insertions in constant-time: FlowBlaze relies on a stash that may be easily filled by an adversary while SilkRoad is limited by both the size of the Bloom Filter and the complexity of the implementation. CHEETAH uses a constant-time stack that is amenable to fast implementation in the dataplane. Existing stateful LBs also suffer from the fact that the 1st-tier of stateless ECMP LBs reshuffle connections to the wrong stateful LB when the number of LBs changes. In contrast, 1st-tier stateless CHEETAH guarantees connections reach the correct stateful LB regardless of changes in the LB pool size.

8 Conclusions

We introduced CHEETAH, a novel building block for load balancers that guarantees PCC and supports any realizable LB mechanisms. We implemented CHEETAH on both software switches and programmable ASIC Tofino switches. We consider this paper as a first step towards unleashing the power of load balancing mechanisms in a resilient manner. We leave the question of whether one can design novel load balancing mechanisms tailored for Layer 4 LBs as well as deployability with existing middleboxes as future work.

Acknowledgements

We would like to thank our shepherd Jitendra Padhye and the anonymous reviewers for their invaluable feedback. The 7th author thanks the participants at the Dagstuhl workshop on "Programmable Network Data Planes". This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 770889). This work was also funded by the Swedish Foundation for Strategic Research (SSF). The work was also supported in parts by a KAW Academy Fellowship and an SSF Framework Grant.

References

[1] Alexa. The top 500 sites on the web. URL: https://www.alexa.com/topsites.

[2] Mohammad Alizadeh, Tom Edsall, Sarang Dharmapurikar, Ramanan Vaidyanathan, Kevin Chu, Andy Fingerhut, Vinh The Lam, Francis Matus, Rong Pan, Navindra Yadav, and George Varghese. Conga: Distributed congestion-aware load balancing for datacenters. *SIGCOMM Comput. Commun. Rev.*, 44(4), August 2014.

[3] Joao Taveira Araujo, Lorenzo Saino, Lennert Buytenhek, and Raul Landa. Balancing on the edge: Transport affinity without network state. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, 2018.

[4] Cheetah authors. Github - cheetah source code, 2020. https://github.com/cheetahlb/. URL: https://github.com/cheetahlb/.

[5] Tom Barbette. Github - FastClick, 2015. https://github.com/tbarbette/fastclick. URL: https://github.com/tbarbette/fastclick.

[6] Barefoot. Tofino: World's Fastest P4 Programmable Ethernet Switch ASIC, September 2019. URL: https://www.barefootnetworks.com/products/brief-tofino/.

[7] Theophilus Benson, Ashok Anand, Aditya Akella, and Ming Zhang. Microte: Fine grained traffic engineering for data centers. In *Proceedings of the Seventh COnference on Emerging Networking EXperiments and Technologies*, CoNEXT '11, 2011.

[8] David Borman, Robert T. Braden, Van Jacobson, and Richard Scheffenegger. TCP Extensions for High Performance. RFC 7323, September 2014. URL: https://rfc-editor.org/rfc/rfc7323.txt, doi:10.17487/RFC7323.

[9] Marco Chiesa, Guy Kindler, and Michael Schapira. Traffic engineering with equal-cost-multipath: An algorithmic perspective. *IEEE/ACM Trans. Netw.*, 25(2):779–792, 2017. URL: https://doi.org/10.1109/TNET.2016.2614247, doi:10.1109/TNET.2016.2614247.

[10] Fabien Duchene and Olivier Bonaventure. Making multipath TCP friendlier to load balancers and anycast. In *25th IEEE International Conference on Network Protocols, ICNP 2017, Toronto, ON, Canada, October 10-13, 2017*, pages 1–10, 2017.

[11] Martin Duke. QUIC-LB: Generating Routable QUIC Connection IDs. Internet-Draft draft-duke-quic-load-balancers-04, Internet Engineering Task Force, May 2019. Work in Progress. URL: https://datatracker.ietf.org/doc/html/draft-duke-quic-load-balancers-04.

[12] Wesley Eddy. TCP SYN Flooding Attacks and Common Mitigations. RFC 4987, August 2007. URL: https://rfc-editor.org/rfc/rfc4987.txt, doi:10.17487/RFC4987.

[13] Daniel E. Eisenbud, Cheng Yi, Carlo Contavalli, Cody Smith, Roman Kononov, Eric Mann-Hielscher, Ardas Cilingiroglu, Bin Cheyney, Wentao Shang, and Jinnah Dylan Hosein. Maglev: A fast and reliable software network load balancer. In *Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation*, NSDI'16, 2016.

[14] Pierre Fraigniaud and Cyril Gavoille. Memory requirement for universal routing schemes. In *Proceedings of the Fourteenth Annual ACM Symposium on Principles of Distributed Computing, Ottawa, Ontario, Canada, August 20-23, 1995*, pages 223–230, 1995.

[15] Rohan Gandhi, Hongqiang Harry Liu, Y. Charlie Hu, Guohan Lu, Jitendra Padhye, Lihua Yuan, and Ming Zhang. Duet: Cloud scale load balancing with hardware and software. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM '14, 2014.

[16] Soudeh Ghorbani, Zibin Yang, P. Brighten Godfrey, Yashar Ganjali, and Amin Firoozshahian. Drill: Micro load balancing for low-latency data center networks. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '17, 2017.

[17] Will Glozer. WRK. URL: https://github.com/wg/wrk.

[18] Keqiang He, Eric Rozner, Kanak Agarwal, Wes Felter, John Carter, and Aditya Akella. Presto: Edge-based load balancing for fast datacenter networks. *SIGCOMM Comput. Commun. Rev.*, 45(4), August 2015.

[19] Christian Hopps. Analysis of an Equal-Cost Multi-Path Algorithm. RFC 2992, November 2000. URL: https://rfc-editor.org/rfc/rfc2992.txt, doi:10.17487/RFC2992.

[20] Christian Hopps. Katran: A high performance layer 4 load balancer, September 2019. URL: https://github.com/facebookincubator/katran.

[21] Intel. Receive-Side Scaling (RSS), 2016. URL: http://www.intel.com/content/dam/support/us/en/documents/network/sb/318483001us2.pdf.

[22] Nanxi Kang, Monia Ghobadi, John Reumann, Alexander Shraer, and Jennifer Rexford. Efficient traffic splitting on commodity switches. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '15, 2015.

[23] David Karger, Eric Lehman, Tom Leighton, Rina Panigrahy, Matthew Levine, and Daniel Lewin. Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the world wide web. In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, STOC '97, 1997.

[24] Naga Katta, Mukesh Hira, Aditi Ghag, Changhoon Kim, Isaac Keslassy, and Jennifer Rexford. Clove: How i learned to stop worrying about the core and love the edge. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, HotNets '16, 2016.

[25] Naga Katta, Mukesh Hira, Changhoon Kim, Anirudh Sivaraman, and Jennifer Rexford. Hula: Scalable load balancing using programmable data planes. In *Proceedings of the Symposium on SDN Research*, SOSR '16, 2016.

[26] Eddie Kohler, Robert Morris, Benjie Chen, John Jannotti, and M. Frans Kaashoek. The click modular router. *ACM Trans. Comput. Syst.*, 18(3):263–297, August 2000. URL: http://doi.acm.org/10.1145/354871.354874, doi:10.1145/354871.354874.

[27] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1-4):157–168, 1968.

[28] Hugo Krawczyk. New hash functions for message authentication. In *Proceedings of the 14th Annual International Conference on Theory and Application of Cryptographic Techniques*, EUROCRYPT'95, 1995.

[29] Linux Foundation. Data plane development kit (DPDK), 2015. http://www.dpdk.org. URL: http://www.dpdk.org.

[30] Marek Majkowski. SYN packet handling in the wild, January 2018. URL: https://blog.cloudflare.com/syn-packet-handling-in-the-wild/.

[31] James McCauley, Aurojit Panda, Arvind Krishnamurthy, and Scott Shenker. Thoughts on load distribution and the role of programmable switches. *SIGCOMM Comput. Commun. Rev.*, 49(1):18–23, February 2019. URL: https://doi.org/10.1145/3314212.3314216, doi:10.1145/3314212.3314216.

[32] Rui Miao, Hongyi Zeng, Changhoon Kim, Jeongkeun Lee, and Minlan Yu. Silkroad: Making stateful layer-4 load balancing fast and cheap using switching asics. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '17, 2017.

[33] Michael David Mitzenmacher. *The Power of Two Choices in Randomized Load Balancing*. PhD thesis, University of California, Berkeley, 1996. AAI9723118.

[34] Felician Nemeth, Marco Chiesa, and Gabor Retvari. Normal forms for match-action programs. In *Proceedings of the 15th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '19, 2019.

[35] NetApplications. Market share for mobile, browsers, operating systems and search engines | NetMarketShare. URL: https://bit.ly/37iRNOK.

[36] NoviFlow. Katran: A high performance layer 4 load balancer, September 2019. URL: https://noviflow.com/noviswitch/.

[37] Vladimir Olteanu, Alexandru Agache, Andrei Voinescu, and Costin Raiciu. Stateless datacenter load-balancing with beamer. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, 2018.

[38] Vladimir Olteanu, Alexandru Agache, Andrei Voinescu, and Costin Raiciu. Stateless datacenter load-balancing with Beamer. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, pages 125–139, Renton, WA, April 2018. USENIX Association. URL: https://www.usenix.org/conference/nsdi18/presentation/olteanu.

[39] Vladimir Andrei Olteanu and Costin Raiciu. Datacenter scale load balancing for multipath transport. In *Proceedings of the ACM SIGCOMM Workshop on Hot topics in Middleboxes and Network Function Virtualization, HotMiddlebox@SIGCOMM 2016, Florianopolis, Brazil, August, 2016*, pages 20–25, 2016.

[40] Rasmus Pagh and Flemming Friche Rodler. Cuckoo hashing. *J. Algorithms*, 51(2), May 2004.

[41] Parveen Patel, Deepak Bansal, Lihua Yuan, Ashwin Murthy, Albert Greenberg, David A. Maltz, Randy Kern, Hemant Kumar, Marios Zikos, Hongyu Wu, Changhoon Kim, and Naveen Karri. Ananta: Cloud scale load balancing. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, SIGCOMM '13, 2013.

[42] Benoît Pit-Claudel, Yoann Desmouceaux, Pierre Pfister, Mark Townsley, and Thomas Clausen. Stateless load-aware load balancing in p4. In *2018 IEEE 26th International Conference on Network Protocols (ICNP)*, pages 418–423. IEEE, 2018.

[43] Salvatore Pontarelli, Roberto Bifulco, Marco Bonola, Carmelo Cascone, Marco Spaziani, Valerio Bruschi, Davide Sanvito, Giuseppe Siracusano, Antonio Capone, Michio Honda, Felipe Huici, and Giuseppe Siracusano. Flowblaze: Stateful packet processing in hardware. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, 2019.

[44] Statcounter. Desktop Operating System Market Share Worldwide. URL: https://gs.statcounter.com/os-market-share/desktop/worldwide.

[45] Statcounter. Desktop vs Mobile vs Tablet Market Share Worldwide. URL: https://gs.statcounter.com/platform-market-share/desktop-mobile-tablet.

[46] Statcounter. Mobile Operating System Market Share Worldwide. URL: https://gs.statcounter.com/os-market-share/mobile/worldwide.

[47] Helen Tabunshchyk. Super Fast Packet Filtering with eBPF and XDP, December 2017. URL: https://bit.ly/2mpoIyO.

[48] Mellanox Technologies. ConnectX®-5 EN Single/Dual-Port Adapter Supporting 100Gb/s Ethernet. URL: https://www.mellanox.com/page/products_dyn?product_family=260&mtag=connectx_5_en_card.

[49] Vladimir P. Chistyakov Valentin F. Kolchin, Boris A. Sevastyanov. *Random allocations*. Washington: Winston, 1978.

[50] John Carl Villanueva. Comparing Load Balancing Algorithms, June 2015. URL: https://www.jscape.com/blog/load-balancing-algorithms.

[51] Weikun Wang and Giuliano Casale. Evaluating weighted round robin load balancing for cloud web services. In *16th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2014, Timisoara, Romania, September 22-25, 2014*, pages 393–400, 2014.

[52] Hong Zhang, Junxue Zhang, Wei Bai, Kai Chen, and Mosharaf Chowdhury. Resilient datacenter load balancing in the wild. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '17, 2017.

[53] Junlan Zhou, Malveeka Tewari, Min Zhu, Abdul Kabbani, Leon Poutievski, Arjun Singh, and Amin Vahdat. Wcmp: Weighted cost multipathing for improved fairness in data centers. In *Proceedings of the Ninth European Conference on Computer Systems*, EuroSys '14, 2014.

APPENDIX

A Proof of Theorem 1

Theorem 1. Given an arbitrarily large number of connections, any load balancer using $O(1)$ memory requires cookies of size $\Omega(\log(k))$ to guarantee PCC under any possible change in the number of active servers, where k is the overall number of servers in the DC that can be assigned to the service with a given VIP.

Proof sketch. We prove the statement of the theorem in the widely adopted Kolmogorov descriptive complexity model [27]. We leverage similar techniques used in the past to demonstrate a variety of memory-related lower bounds for shortest-path routing problems [14].

Let R be the set of all the possible connection identifiers. Let C be the set of all possible cookies. Let $S = \{s_1, \ldots, s_k\}$ be the set of servers. We assume $|R| \gg |S|$, which is the most interesting case in real-world datacenters. Suppose, by contradiction, that there exists an LB which uses $O(1)$ memory with cookies of size smaller than $\log(k)$ bits that guarantees under any arbtirary number of changes in the subset of active servers $A \subseteq S$. For any possible set of active servers A, the LB maps a new incoming connection identifier $r \in R$ to a certain server $s \in S$, *i.e.*, the LB logic maps incoming connections using an arbitrarily function $f : R \times 2^A \to S$.

Let us now restrict our focus to the $|S|$ distinct sets of active servers in which only a single server is active, *i.e.*, $A_1 = \{s_1\}, \ldots, A_k = \{s_k\}$. Depending on the time instance when a connection $r \in R$ arrives, the connection may be mapped to any of these servers. The mapping must be preserved for the entire duration of the connection, which means the LB must be able to forward any future packet belonging to r regardless of the current set of active servers. Therefore, the LB must be able to distinguish among $|R| \times |S|$ distinct possible mappings between connections and servers. Consider our cookie with l bits, where $l < |S| = \log(k)$. This information allows us to distinguish among $|R| \times l$ possible mappings, which leaves $|R| \times (|S| - l) = O(|R||S|)$ mappings to the LB memory. This is a contradiction since we assumed the LB uses a memory of $O(1)$. □

It is trivial to verify that the above theorem holds even if one wants to implement an advanced LB mechanism, *e.g.* round-robin, least-loaded, even without allowing any changes in the number of servers.

B Constant-size cookies

Minimizing PCC violations with constant-size cookies: keeping a history at the LB. A simple way to deal with changes in the number of active servers when using a uniform-hash load balancing mechanism is to encode in the cookie a *tag* that can be used by the LB to uniquely identify a previous configuration of active servers. The LB stores the last n configurations of active servers and identifies them using $\log_2(n)$ bits, which are encoded into the cookie. Thus, as long as a connection is mapped to a server that existed in any of the last n configurations, the connection will be unbroken. The LB can modify the cookie associated with a connection to use "fresher" cookies. [8]. Clearly, if an operator drains a server, *i.e.*, purposely does not assign new connections to it, any remaining connections will be broken after n changes in the tag.

Minimizing load imbalances using constant-size cookies: adding a hash function index to the cookie. A common technique to reduce load imbalances in a system is to rely on "power-of-two" choices mechanisms to map a request to a server. When the LB receives a new incoming connection, it computes the hash of the connection identifier using two different hash functions h_1 and h_2, whose outputs are used to select two servers. The LB then compares the load of the two selected servers and maps the incoming connection to the least loaded server among these two (according to some definition of "load"). The main issue in the DC context with power-of-two choice load balancing mechanisms is that the LB needs to remember for each connection, which of the two servers was used to serve the connection, that is, the LB must be stateful. CHEETAH can support a power-of-two choices LB mechanism simply by storing in the cookie a single bit that identifies which of the two hash functions is to be used for that connection. Similarly, multiple hashes could be supported by increasing the size of the cookie logarithmically in the number of hash functions. We note that when the number of servers change, connections may break. We refer the reader to the previous paragraph on how to deal with PCC violations with constant-size cookies.

C TCP timestamp encoding

We decided to encode the CHEETAH cookie into the 16 most significant bits of the TCP timestamp. We acknowledge that alternative ways to encode information into the TCP timestamps are possible [39] but require modifications to the servers.

Encoding cookies in the TCP timestamp option. TCP end-hosts use TCP timestamp options to estimate the RTT of the connection. The timestamp consists of a 64-bit pair (TS_{val}, TS_{ecr}), where TS_{val} and TS_{ecr} are the 32-bit timestamps set by the sender and receiver of the packet, respectively. When an end-host receives a packet, it echoes the TS_{val} back to the other end-host by copying it into the TS_{ecr}. We leverage the TS_{val} to carry the CHEETAH cookie on every packet directed towards the client, which will echo it back as the TS_{ecr}. We encode the cookie in the 16 most significant bits of the 32-bit TS_{val} for every packet directed towards a client.[9] We must therefore fix the original 16 most significant bit of the TS_{ecr} before it is processed by the TCP stack of the server. This can be done in the load-balancer or on the server itself. Our measurements of the top 100 ranked Alexa websites [1] reported in App. D shows that the minimal unit of a timestamp is 1 millisecond. This means that the least significant 16 bits of the timestamp would wrap up every 2^{16}ms.

TCP timestamps are mostly supported in today's OSes. We ran a small experiment to verify whether today's client devices support the echoing of TCP timestamp options back to the servers. We tested the latest OSes available in both recent smartphones and desktop PCs: Google Android 9, iOS 13, Ubuntu 18.04, Microsoft Windows 10, and MacOS 10.14. We observed that all except Microsoft Windows correctly negotiate and echo the TCP timestamp option when the server requires to use it. Based on some recent measurements, more than 98% of the smartphone and tablet devices are either using Android or iOS [46]. Smartphone devices are the most common type of devices, representing 53% of all devices [45]. For desktop devices, Windows is the predominant OS with over 75% of the desktop share whereas MacOS represent a 16% of this share [44]. For Windows desktop devices, a cloud operator can either encode the cookie in the QUIC header (69% of the Windows users use Google Chrome, compatible with QUIC [35]), IPv6 address, or install stateful information into the LB for these devices.

C.1 Fixing the timestamp in the load balancer

For each server, we keep in memory two versions of the 16 most significant bits (MSB) replaced by our cookie: the current one and the previous one. We use one bit of the cookie to remember the version of the original MSB for every given packet sent to the client. When a wrap up of the timestamp happens, we set the oldest MSB bucket to the new MSB timestamp of the server, and we change the version bit of outgoing packets to designate that one. When a packet is received, the cookie is read, then the original MSB given by the version bit found in the packet reader is rewritten back in

[8] This operation requires a bit more than one Round-Trip-Time (RTT) to update the client from the LB (the packet must first be processed by the server, which has to send an acknowledgment to the client with the updated cookie

[9] Timestamp options have already been used in the past for protecting against SYN flood attacks, *e.g.*, TCP SYN cookies [12]. We note that several large cloud networks completely disable TCP SYN cookies and rely on different mechanisms to handle SYN flood attacks [30], thus CHEETAH would not cause deployment issues. We note that an alternative implementation of CHEETAH, in which the mapping with the server is only performed at the end of the 3-way handshake does not prevent the cloud provider from using TCP SYN cookies. This solution requires all the servers to use exactly the same parameters when generating the TCP SYN cookies and is left as a minor future work extension.

the timestamp of the packet. To avoid having clients sending very old cookies, we rely on TCP keepalives set to 25 seconds, which allows us to both detect timestamp wraps at the LB[10] and deal with TCP PAWS packet filters[11].

C.2 Fixing the timestamp on the server

We modified the Linux Kernel 5.1.5 (available at [12]) to enable the same mechanism directly on the server. Doing so increase the server packet processing time by only 0.2%, as the timestamp is parsed in any cases on the server. This enables the server to keep the value of the cookie, and directly encode the cookie in the TS_{val}. Such returning packet do not need to go through the load balancer and allows the use of DSR. Having the cookie on both sides of the load-balancer is also needed for the stateful implementation.

D Alexa Top100 Timestamp Measurements

We ran a comprehensive set of measurements to determine the granularity of the TCP timestamp unit utilized by the largest service providers according to the Alexa Top100 ranking [1]. We downloaded large files from each the top 15 ranked web sites and extracted both the TCP timestamp TS_{val} options and the client side timestamp. We then computed the difference between the TCP last and first timestamps and divided this amount by the different between the client measured last and first (non-TCP) timestamps. The result is the granularity of the server-side TCP timestamp unit. We report the results in Table 2. All the service providers using TCP timestamps have a granularity of at least 1ms. This means the timestamp wraps every $2^{16} \approx 65$ seconds when using CHEETAH to support these services.

Table 2: Measured TCP timestamp granularity for different websited. Some service providers do not use TCP timestamp options.

Web site	**TS granularity**	**Method**
drive.google.com	1ms	gdown
dropbox.com	1ms	wget
twitch.tv	1ms	watch video
weobo.com	1ms	watch video
bilibili.com	N.A.	-
pan.baidu.com	N.A.	-
reddit.com	4ms	watch video
qq.com	4ms	watch video
instagram.com	4ms	watch video
onedrive.live.com	N.A.	-
facebook.com	1ms	watch video
twitter.com	N.A.	-
imdb.com	10ms	watch video

[10] Detecting a wrap at the server is a straightforward operation. To detect a wrap at the LB, we need to guarantee the server sends a packet at least once every 2^{15}ms$\approx$ 33s.This would typically be the case for every real-world Internet service.

[11] We note that flipping the MSB of the timestamp (*i.e.* the version bit) every time a wrap is detected may create problems with PAWS [8]. PAWS is a TCP mechanism that discards packets with TCP timestamps that are not “fresh”, *i.e.*, a timestamp is considered *old* if the difference between the latest received timestamp and the newest received timestamp is smaller than 2_{31}. To avoid enabling this condition, it is sufficient to guarantee the server keeps a TCP keep-alive timer of $(2^{15} - \max RTT)$ milliseconds. Assuming a $\max RTT$ of 5 seconds, we set the keep-alive to a conservative value of 25 seconds. With 100K connections, the bandwidth overhead is just 0.00002% on a 100Gbps server interface.

[12] https://github.com/cheetahlb/linux