Network-Accelerated Consensus for Read-Intensive Workloads

Extended version of our NSDI '20 [1] and IEEE/ACM Transaction on Networking [2] papers

Hatem Takruri, Ibrahim Kettaneh, Ahmed Alquraan, Samer Al-Kiswany
University of Waterloo, Canada

Abstract

We present FLAIR, a novel approach for accelerating read operations in leader-based consensus protocols. FLAIR leverages the capabilities of the new generation of programmable switches to serve reads from follower replicas without compromising consistency. The core of the new approach is a packet-processing pipeline that can track client requests and system replies, identify consistent replicas, and at line speed, forward read requests to replicas that can serve the read without sacrificing linearizability. An additional benefit of FLAIR is that it facilitates devising novel consistency-aware load balancing techniques.

Following the new approach, we designed FlairKV, a key-value store atop Raft. FlairKV implements the processing pipeline using the P4 programming language. We evaluate the benefits of the proposed approach and compare it to previous approaches using a cluster with a Barefoot Tofino switch. Our evaluation indicates that, compared to state-of-the-art alternatives, the proposed approach can bring significant performance gains: up to 42% higher throughput and 35-97% lower latency for most workloads.

1. Introduction

Replication is the main reliability technique for many modern cloud services [3, 4, 5] that process billions of requests each day [5, 6, 7]. Unfortunately, modern strongly-consistent replication protocols [8] – such as multi-Paxos [9], Raft [10], Zab [11], and Viewstamped replication (VR) [12] – deliver poor read performance. This is because these protocols are leader-based: a single leader replica (or leader, for short) processes every read and write request, while follower replicas (followers for short) are used for reliability only.

Optimizing read performance is clearly important; for instance, the read-to-write ratio is 380:1 in Google's F1 advertising system [13], 500:1 in Facebook's TAO [7], and 30:1 in Facebook memcached deployments [14]. Previous efforts have attempted to accelerate reads by giving read leases [15] to some [16] or all followers [3, 17, 18], While holding a lease, a follower can serve read requests without consulting the leader; each lease has an expiration period. Unfortunately, this approach complicates the system's design, as it requires careful management of leases, affects the write operation – as all granted leases need to be revoked before an object can be modified – and imposes long delays when a follower holding a lease fails [3, 16].

Alternatively, many systems support a relaxed consistency model (e.g., eventual [4, 19, 20, 21, 22, 23] or read-your-write [7, 23, 24]), in exchange for the ability to read from followers, albeit the possibility of reading stale data.

In this paper, we present the fast, linearizable, network-accelerated client reads (FLAIR), a novel protocol to serve reads from follower replicas with minimal changes to current leader-based consensus protocols without using leases, all while preserving linearizability. In addition to improving read performance, FLAIR improves write performance by reducing the number of requests that must be handled by the leader and employing consistency-aware load-balancing.

FLAIR is positioned as a shim layer on top of a leader-based protocol (§3). FLAIR assumes a few properties of the underlying consensus protocol: the operations are stored in a replicated log; at any time, there is at most one leader in the system that can commit new entries in the log; reads served by the leader are linearizable; and after committing an entry in the log, the leader knows which followers have a log consistent with its log up to that entry. These properties hold for all major leader-based protocols (Raft [10], VR [12], DARE [25], Zookeeper [4], and multi-Paxos [26, 27, 28]).

FLAIR leverages the power and flexibility of the new generation of programmable switches. The core of FLAIR is a packet-processing pipeline (§4) that maintains compact information about all objects stored in the system. FLAIR tracks every write request and the corresponding system reply to identify which objects are stable (i.e., not being modified) and which followers hold a consistent value for each object, then uses this information to forward reads of stable objects to consistent followers. Followers optimistically serve reads and the FLAIR switch validates read replies to detect stale values. If the switch suspects that a reply from a follower is stale, it will drop the reply and resubmit the read request to the leader.

An additional benefit of FLAIR is that it facilitates the building of novel consistency-aware load balancing techniques. In systems that grant a lease to followers [3, 16, 17, 18], clients send read requests to a randomly selected follower. If the follower does not hold a lease, it blocks the request until it obtains a lease, or it forwards the request to the leader; either way, this approach adds additional delay. FLAIR does not incur this inefficiency as FLAIR load balances read requests only among followers that hold a consistent value for the requested object. In this paper we design three consistency-aware load balancing techniques (§6): random, leader avoidance, and load awareness.

Unlike other systems that use switch's new capabilities [29, 30, 31], FLAIR does not rely on the controller to update the switch information after every write operation, as this approach would add unacceptable delays. Instead, FLAIR piggybacks control messages on system replies, and the switch extracts and processes them.

Despite its simplicity, implementing this approach is complicated by the limitations of programmable switches (§2) and the complexity of handling switch failures, network partitioning, and packet loss and reordering (§4).

To demonstrate the powerful capabilities of the proposed approach, we prototyped FlairKV (§6), a key-value store built atop Raft [10]. We made only minor changes to Raft's implementation [32] to enable followers to serve reads, make the leader order write requests following the sequence numbers assigned by the switch, and expose leader's log information to the FLAIR layer. The packet-processing pipeline was implemented using the P4 programming language [33]. We implemented the three aforementioned load-balancing techniques (§6).

Our evaluation of FlairKV (§7) on a cluster with a Barefoot Tofino switch shows that FLAIR can bring sizable performance gains without increasing the complexity of the leader-based protocols or the write operation overhead. Our evaluation with different read-to-write ratios and workload skewness shows that FlairKV brings up to 2.8 times higher throughput than an optimized Raft implementation, at least 4 times higher throughput compared to Viewstamped replication, Raft and FastPaxos, and up to 42% higher throughput and up to 35-97% lower latency for most workloads compared to state-of-the-art leases-based design [3, 18].

The performance and programmability of the new generation of switches opens the door for the switches to be used beyond traditional network functionalities. We hope our experience will inform a new generation of distributed systems that co-design network protocols with systems operations.

2. Background

In this section, we present an overview of leader-based consensus protocols, followed by a look at the new programmable switches and their limitations.

2.1. Leader-based Consensus

Leader-based consensus (LC) protocols [10, 11, 12, 25, 26, 34] are widely adopted in modern systems [4, 5, 6, 18]. The idea of having a leader that can commit an operation in a single round trip dates back to the early consensus protocols [9, 35]. Having a leader reduces contention and the number of messages, which greatly improves performance [9, 26].

LC protocols divide time into terms (a.k.a. views or epochs). Each term has a single leader; if the leader fails, a new term starts and a new leader is elected.

Clients send write requests to the leader (1 in Figure 1). The leader appends the request to its local log (2) and then sends the request to all follower replicas (3). A follower appends the request to its log (4) before sending an acknowledgment to the leader (5). If the leader receives an acknowledgment from a majority of its followers, the operation is considered *committed*. The leader *applies* the operation to its local state machine (e.g., in memory key-value store in Figure 1) in (6), then acknowledges the operation to the client (7). The leader will asynchronously inform the followers that it committed the operation. Followers maintain a *commit_index*, a log index pointing to the last committed operation in the log; when a follower receives the commit notification, it advances its *commit_index* and applies the write to its local store.

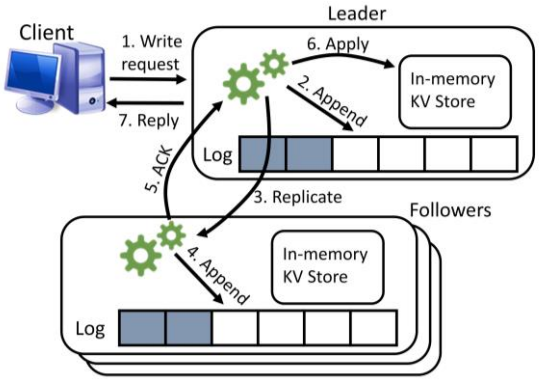


Figure 1. The path for a write operation.

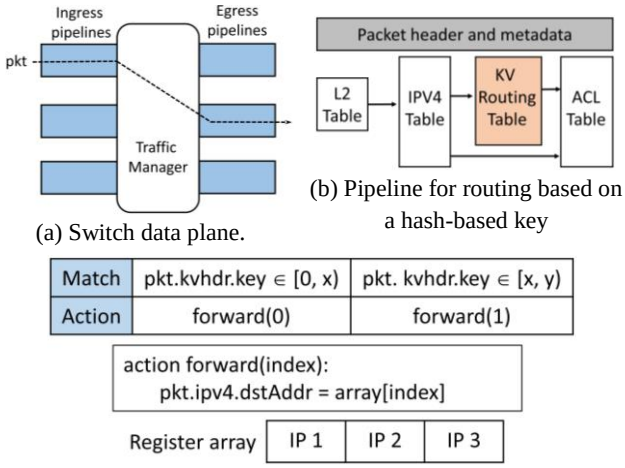
The replicated log has two properties that make it easy to reason about: it is guaranteed that if an operation at index *i* is committed, then every operation with an index smaller than *i* is committed as well; and if a follower accepts a new entry to its log, it is guaranteed that its log is identical to the leader's log up to that entry.

Client read requests are also sent to the leader. In Raft, the leader sends a heartbeat to all followers to make sure it is still the leader. If a majority of followers reply, the leader serves the read form its local store: it will check that all committed operations related to the requested object are applied to the local store before serving the request.

A common optimization is the leader lease optimization. Instead of collecting a majority of heartbeats for every read request, a majority of the followers can give the leader a lease [10, 26]. While holding a lease, the leader serves reads locally without contacting followers. Unfortunately, even with this optimization, the performance of the leader-based protocols is limited to a single-node performance.

2.2. Programmable Switches

Programmable switches allow the implementation of an application-specific packet-processing pipeline that is deployed on network devices and executed at line speed. A number of vendors produce network-programmable ASICs, including Barefoot's Tofino [36], Cavium's XPliant [37], and Broadcom Trident 3 [38].



(a) Switch data plane.

(b) Pipeline for routing based on a hash-based key

(c) Simple match-action stage for routing based on a hash-based key for the KV routing table in subfigure (b)

Figure 4. Switch data plane.

Figure 4(a) illustrates the basic data plane architecture of modern programmable switches. The data plane contains three main components: ingress pipelines, a traffic manager, and egress pipelines. A packet is first processed by an ingress pipeline before it is forwarded by the traffic manager to the egress pipeline that will finally emit the packet.

Each pipeline is composed of multiple stages. At each stage, one or more tables match fields in the packet header or metadata; if a packet matches, the corresponding action is executed. Programmers can define custom headers and metadata as well as custom actions. Each stage has its own dedicated resources, including tables and register arrays (a memory buffer). Figure 4(b) shows a simple example of a pipeline that routes a request to a key-value store based on the key, and Figure 4(c) shows the details of the KV routing stage. The stage forwards the request based on the key in the packet's custom L4 header. The programmer implements a `forward()` action that accesses the register array holding nodes' IP addresses. An external controller can modify the register array and the table entries.

Stages can share data through the packet header and small per-packet metadata (a few hundred bytes in size) that is propagated between the stages as the packet is processed throughout the pipeline (Figure 4(b)). The processing of packets can be viewed as a graph of match-action stages.

Programmers use domain-specific languages like P4 [39] to define their own packet headers, define tables, implement custom actions, and configure the processing graphs.

Challenges. While programmable ASICs and their domain-specific languages significantly increase the flexibility of network switches, the need to execute custom actions at line speed restricts what can be done. To process packets at line speed, P4 and modern programmable ASICs have to meet strict resource and timing requirements. Consequently, modern ASICs limit (1) the number of stages per pipeline, (2) the number of tables and registers per stage, (3) the number of times any register can be accessed per packet, (4) the amount of data that can be read/written per-packet per register, (5) the size of per-packet metadata that is passed between stages. Finally, modern ASIC's lack support of loops or recursion.

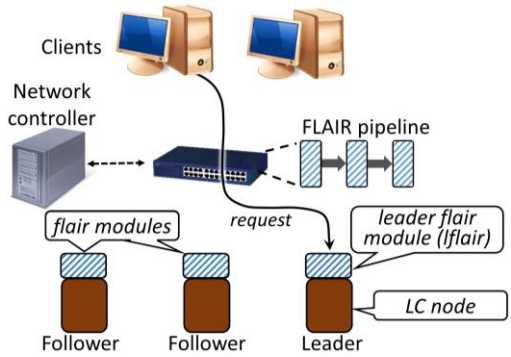


Figure 2. System architecture. The solid arrow shows a client request, while the dashed arrow show control messages.

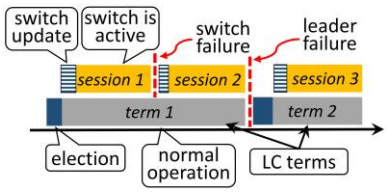


Figure 3. FLAIR sessions. Time is divided into terms. Each term starts with a leader election. Each term has one or more sessions that start with updating the switch data.

3. FLAIR Overview

FLAIR is a novel protocol that targets deployments in a single data center. Figure 2 shows the system architecture, which consists of a programmable switch, a central controller, and storage nodes. Typically, multiple FLAIR instances are deployed with each serving a disjoint set of objects. For simplicity, we present a FLAIR deployment with one replica set (i.e., one leader and its followers).

FLAIR is based on the following assumptions; the network is unreliable and asynchronous, as there are no guarantees that packets will be received in a timely manner or even delivered at all, and there is no limit on the time a node or switch takes to process a packet. Finally, FLAIR assumes a non-byzantine failure model in which nodes and switches may stop working but will never send erroneous messages.

FLAIR divides time into sessions (Figure 3). During a session the leader is bonded to a single switch that runs the FLAIR pipeline. Each session has a unique id that is assigned in a strictly increasing order. A session ends when a leader fails or the leader suspects that the switch has failed. An LC term may have one or more sessions, but a session does not span multiple terms.

A session starts with the FLAIR module at the leader (dubbed the *lflair* module) incrementing the session id, committing it to the LC log, updating the switch information about the objects in the system, then activating the session at the switch. *lflair* module keeps the switch's information up

to date while in an active session. If the switch does not have an active session it drops all FLAIR packets.

Clients. FLAIR is accessed through a client library with a simple read/write/delete interface. Read (get) and write (put) read or write entire objects. The library adds a special FLAIR packet header to every request, that contains an operation code (e.g., read) and a key (a hash-based object identifier).

Controller. Our design targets data centers that use a SDN network following a variant of the multi-rooted tree topology [40, 41]. A central controller uses OpenFlow [42] to manage the network by installing per-flow forwarding, filtering, and rewriting rules in switches.

As with previous projects that leverage SDN capabilities [29, 31, 43, 44], the controller installs forwarding rules to guarantee that every client request for a range of keys served by a single replica set is passed through a specific switch (dubbed FLAIR switch); that switch will run the FLAIR logic for that range of keys. The controller typically selects a common ancestor switch of all replicas and installs rules to forward system replies through the same switch. Only client request/replies are routed through the FLAIR switch, leader-follower messages do not have the FLAIR header nor are necessarily routed through the FLAIR switch.

While this approach may create a longer path than traditional forwarding, the effect of this change is minimal. Li et al. [43] reported that for 88% of cases, there is no additional latency, and the 99th percentile had less than 5 μs of added latency. This minimal added latency is due to the fact that the selected switch is the common ancestor of target replicas and client packets have to traverse that switch anyway.

On a switch failure, the controller selects a new switch and updates all the forwarding rules accordingly. The controller load balances the work across switches by assigning different replica sets to different switches.

Storage Nodes. The storage nodes run the FLAIR and LC protocols. For read requests, before serving a read, followers verify that all committed writes to the requested object have been applied to the follower's local storage.

Write requests are processed by the leader. After a successful write operation, the leader passes to the *lflair* module the log index at which the write was committed and the list of followers that accepted the write operation and have a consistent log up to that log index. The *lflair* encodes this list into a compact bitmap and uploads it and the log index to the switch (piggybacked on the write reply).

Programmable Switch. The switch is a core component of FLAIR: it tracks every write request and the corresponding reply to identify which objects are stable (not being modified) and which replicas have a consistent value of each object (encoded in the bitmap provided by the *lfair* module). If a read is issued while there are outstanding writes for the target object (i.e., writes without corresponding replies), the read is forwarded to the leader. If a read request is processed by the switch when there are no outstanding writes to the requested object, the switch forwards the request to one of the followers included in the last bitmap for the object sent by the *lflair* module. Followers optimistically serve read requests. The switch inspects every read reply; if it suspects that a follower returned stale data (Section 4.4), it will conservatively drop the reply and forward the request to the leader. FLAIR forwards all writes to the leader.

FLAIR also includes techniques to handle multiple concurrent writes to the same object (Section 4.3), packets reordering (Section 4.6), and tolerating switch, node, and network failures (Section 4.6).

4. System Design

4.1. Network Protocol

Packet format. FLAIR introduces an application-layer protocol embedded in the L4 payload of packets. Similar to many other storage systems [29, 31, 43], FLAIR uses UDP to issue client requests in order to achieve low latency and simplify request routing. Communication between replicas uses TCP for its reliability. A special UDP port is reserved to distinguish FLAIR packets; for UDP packets with this port, the switch invokes the FLAIR custom processing pipeline. Other switches do not need to understand the FLAIR header and will treat FLAIR packets as normal packets. In this way, FLAIR can coexist with other network protocols.

Figure 5 shows the main fields in the FLAIR header. We briefly discuss the fields here (a detailed discussion of the protocol is presented next):

- OP: the request type. Clients populate this field in the request packet (e.g., read, or write); replicas populate this field in the reply packets (e.g., read_reply, write_reply).
- KEY: hash-based object identifier.
- SEQ: a sequence number added by the switch. The switch increments the sequence number on every write operation.
- SID: a unique session id. The <SID, SEQ> combination represents a unique identifier for every write request.
- LOG_IDX: a log index. In a write_reply, the log index indicates the index at which the write was committed. For reads, the switch populates LOG_IDX to make sure the followers' logs are committed and applied up to that index.
- CFLWRS: In write_reply, the CFLWRS is a map of the followers that have a consistent log up to LOG_IDX.

Following the FLAIR header is the original LC protocol payload, which includes the value for read/write operations.

4.2. Switch Data Structures

To process a read request, the switch performs two specific tasks (Section 4.4). First, it forwards read requests to consistent followers while balancing the load among them. Second, it verifies the read replies to preserve safety. To perform these tasks, the switch maintains two data structures: a session array and a key group array.

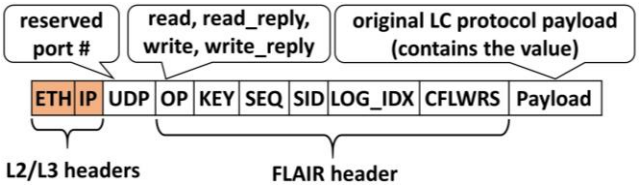


Figure 5. FLAIR packet format.

```
SessionArrayEntry {
  bit<1>  is_active;
  bit<32> session_id;
  bit<32> leader_ip;
  bit<64> session_seq_num;
  bit<48> heartbeat_tstamp; }
```

```
KGroupArrayEntry {
  bit<1>  is_stable;
  bit<64> seq_num;
  bit<64> log_idx;
  bit<8>  consistent_followers;
}
```

Listing 1. Session and kgroup entries. The numbers indicate the field size in bits.

Session array. A single switch typically supports multiple replica sets (i.e., FLAIR+LC instances) with each set storing a disjoint set of keys. Each entry in the session array maintains the session status for a single replica set. An entry contains an is_active flag, session id, leader IP address, current session sequence number, and the timestamp of the last heartbeat received from the *lflair* module (Listing 1). When is_active is true, we say the session is *active*, which indicates that the session entry and kgroup array are consistent with the leader's information. The switch processes packets using the FLAIR custom pipeline only if the session is active; otherwise, it will drop all FLAIR packets, rendering the system unavailable to clients until the switch can reach the *lflair* module and sync its session entry and key group array.

Key group (KGroup) array. To decide if followers can serve a certain read request, the switch needs to maintain information about which followers have the latest committed value of every object. Maintaining such information in the switch ASIC's memory is not feasible; instead, FLAIR groups objects based on their key and maintains aggregate information per group. We use the most significant *k* bits of the key to map an object to a key group (kgroup).

Every FLAIR+LC instance has a dedicated kgroup array. Each entry in the array (Listing 1) contains the status of a single kgroup, including an is_stable flag that indicates if all objects in the kgroup are stable. If a kgroup is not stable (is_stable is false), this indicates that at least one object in the kgroup is being modified (i.e., has an outstanding write in the system). The array entry also includes the sequence number (seq_num) of the last write request processed by the switch for any object in the kgroup, the log index (log_idx) of the last successful write to any object in the kgroup, and the consistent_followers bitmap pointing to all followers that have a consistent log up to log_idx.

4.3. Handling Write Requests

To issue a write request, a client populates the OP and KEY fields of the FLAIR packet header and puts the value in the payload, then sends the request.

When the switch receives the request, it will mark the corresponding kgroup entry as unstable. The switch will increment the session_seq_num in the session array and use it to populate the sequence number (seq_num) in the kgroup entry and the sequence number (SEQ) in the request header. Finally, the switch populates the session id (SID) field in the header and forwards the request to the leader.

The *lflair* module will verify that the session id is valid, and will pass the write request to the leader. The leader verifies that the <SID, SEQ> combination is larger than the <SID, SEQ> number of any previous write request it ever received, else it will drop the packet. The LC leader will process the write request following the LC protocol (Section 2.1): it will replicate the request to all followers, and when a majority of followers acknowledge the operation, the write operation is considered committed. A follower will acknowledge a write operation only if its log is identical to the leader's log up to that entry.

For the write reply, the leader will pass the following to the *lflair* module: the LC protocol payload for the write_reply, the log index at which the write was committed, and the list of followers that acknowledged the write. The *lflair* module will create the write reply packet with the leader provided payload, and will populate the LOG_IDX and the bitmap of the consistent followers (CFLWRS) using the information provided by the leader. *lflair* module populates the sequence number (SEQ) in the write_reply header using the SEQ of the corresponding write request. The *lflair* module then sends the write_reply packet.

The switch will process the write_reply header and verify its session id. The switch will compare the sequence number (SEQ) of the reply to the sequence number (seq_num) in the kgroup entry; if they are equal, this signifies that no other write is concurrently being processed in the system for any object in the kgroup. Consequently, it will update the log_idx and the consistent_followers fields in the kgroup entry using the values in the write reply. Then it will mark the kgroup stable and forward the reply to the client.

If the sequence number in the reply is smaller than the sequence number in the kgroup entry, this indicates that a later write to an object in the same kgroup has been processed by the switch. In this case, the switch forwards the write reply to the client without modifying the kgroup entry. The kgroup entry remains unstable until the last write to the kgroup (with a SEQ number in the write_reply equal to the seq_num in the kgroup entry) is acknowledged by the leader.

In a nutshell, the switch acts as a look-through metadata cache. Write requests invalidate the switch metadata related to the accessed kgroup, and write replies update the kgroup metadata at the switch. As we see next, the kgroup metadata is used to consistently load balance reads.

4.4. Handling Read Requests

Clients fill the OP and KEY fields of the FLAIR header and send the request. When the switch receives the request, it will check the kgroup entry. If the entry is stable, the switch will fill the sequence number (SEQ) and log index (LOG_IDX) header fields using the values in the kgroup entry. Then it will forward the request to one of the followers indicated in the consistent_followers bitmap. Section 6.2 details our load balancing techniques.

If the kgroup entry is not stable, the switch forwards the read request to the leader. We note that there is a chance for false positives in this design, as a single write will render all the objects in the same kgroup unstable. This is a drawback of maintaining information per group of keys. This inefficiency is incurred by leases-based protocols as well, as they maintain a lease per group of objects.

When a follower receives a read request, the follower's FLAIR module validates the request, then calls advance_then_read(LOG_IDX, key) routine, which compares the follower's commit_index to LOG_IDX. If the commit_index is smaller, the follower advances its commit_index to equal LOG_IDX, apply all the log entries to the local store, then serve the read request. The FLAIR module will populate the read_reply header; for the SEQ and SID fields, it will use the values found in the read request header.

We note that it is safe to advance the follower's commit_index to match the LOG_IDX in the read request, as the switch forwards read requests to a follower only if the leader indicates that all entries in the log up to that log index are committed, and that this specific follower is one of the replicas that have a log consistent to the leader's log up to that index. We discuss FLAIR correctness in Section 5.

When the switch receives a read_reply from a follower, it validates the session id, then verifies that the SEQ number of the read_reply equals the seq_num of the kgroup entry. If the sequence numbers are not equal, this signifies that a later write request was processed by the switch and there is a chance the follower has returned stale value. In this case, the switch drops the read_reply, generates a new read request using the KEY field from read_reply packet, and submits the read request to the leader. If the sequence number of the read_reply equals the sequance number in the kgroup entry, the switch forwards the reply to the client.

If a read request is forwarded to the leader, the *lflair* module verifies the session id, then calls advance_then_read(LOG_IDX, key). The switch verifies that the leader reply is valid (i.e., has the correct session id) before forwarding it to the client, without checking the seq_num in the kgroup entry.

4.5. Session Start Process

On the start of a new session, the *lflair* module reads the last session id from the LC log, increments it, and commits the new session id to the LC log. Then the *lflair* module asks the central controller for a new switch. The central controller neutralizes the old switch (making it drop all FLAIR packets) and reroutes FLAIR packets to a new switch, then confirms the switch change to the *lflair* module. THis step guarantees that at any time at most one FLAIR switch is active. The *lflair* module updates the session entry (Listing 1) at the switch with the current leader IP and session id. For each new session, session_seq_num is reset to zero.

Populating the kgroup array. The *lflair* module maintains a copy of the kgroup array similar to the one maintained by the switch. If the leader did not change between sessions (e.g., the session change is due to switch failure), the kgroup array at the *lflair* module is up to date. The *lflair* module will set the seq_num entry in all kgroup entries to zero (equal to the session_seq_num in the session entry)., and upload it to the switch.

If the kgroup array at the *lflair* module is empty – for instance, after electing a new leader – the *lflair* module will query the leader for three pieces of information: its commit_index, the list of followers with the same commit_index, and a list of all uncommitted operations in the log (i.e., the operations after the commit_index in the log). The list of uncommitted operations is typically small, as it only includes operations that were received before the end of the last term but were not committed yet. The *lflair* module will traverse the list of uncommitted writes and mark their target kgroup entries unstable. For all other kgroup entries, the *lflair* module will mark them stable and set their seq_num to zero, log_idx to the leader's commit_index, and consistent_followers to include all the followers that have the same commit_index as the leader's. After updating the session entry and the kgroup array at the switch, the *lflair* module activates the switch session (sets is_active to true).

4.6. Fault Tolerance

Follower Failure. We rely on the LC protocol to handle follower failures. To avoid sending read requests to a failing follower, the leader notifies the *lflair* module when it detects the failure of a follower. The *lflair* module removes the follower from the switch-forwarding table (Section 3).

Leader Failure. On leader failure, a new leader is elected and a new term starts. The new leader informs the *lflair* module of the term change; and the *lflair* module starts a new session (Section 3).

The *lflair* module sends periodic heartbeats to the switch. Upon receiving a heartbeat, the switch determines whether it is from the current session. If the heartbeat is valid, the switch updates the heartbeat_timestamp in the session array and replies to the *lflair* module.

Switch Failure. If the *lflair* module misses the switch heartbeats for a *switch_stepdown* period of time (3 heartbeats in our prototype), the *lflair* module will suspect that the switch has failed and will start a new session. For efficiency (i.e.

does not affect safety), if the switch misses three heartbeats from the leader, it will deactivate the session.

Network Partitioning. If a network partition isolates the switch from the leader, the leader treats it as a failed switch, as detailed above. If a network partition isolates the switch from a follower, read requests forwarded to the follower will time out and the client will resubmit the request. This failure affects performance, but not correctness. Upon determining that a follower is not reachable, the leader removes it from the forwarding table, as in the case of the failed follower described above.

Packet Loss. If a read or write request is lost, the client times out and resubmits the request. If a write reply is lost before reaching the switch, the kgroup entry will remain unstable until a new write operation to any key in the kgroup succeeds. While the kgroup entry is not stable, all read requests are forwarded to the leader.

Packet Reordering. It is critical for FLAIR correctness that the leader processes write requests in the same order that they are processed by the switch. Every write operation gets a unique <SID, SEQ> number. The switch marks a kgroup entry unstable until the leader replies to the last write issued for a key in the kgroup. Consequently, if the leader processes the requests out of order, the switch will incorrectly mark a kgroup stable while the out-of-order writes are modifying its objects. To prevent this scenario, the leader keeps track of the largest <SID, SEQ> it has ever processed and drops any write request with a smaller number. While session numbers (SIDs) are maintained in the log, the largest processed sequence number is retained in memory. If the leader fails, the new leader starts a new session, increments the session id (SID), and sets the session sequence number (SEQ) to zero.

5. Correctness

FLAIR guarantees linearizability, which means that concurrent operations must appear to be executed by a single machine. FLAIR relies on the LC protocol for any operation that updates the log and for reads from the leader.

FLAIR only adds the ability to serve reads from followers. In this section, we sketch out the proof of FLAIR correctness when the read is served by a follower. Further, we used the TLA+ model checking tool [45] to verify the FLAIR correctness. We started from Raft's TLA+ specification [46] and extended it with a formal specification for our protocol and new invariants to validate the linearizability of reads. Appendix B presents the TLA+ specification.

Safety. FLAIR guarantees that all read replies are linearizable. FLAIR trusts that the leader's read replies are linearizable and forwards them to the client. For reads served by followers, FLAIR guarantees that the read reply returns an identical value, as if the read was served by the leader. This is guaranteed using the following two steps:

First, when the switch receives a read request, the switch forwards that request to followers only when the switch has an active session and the kgroup entry is stable. This signifies that the switch information is up-to-date with the *lflair* module's information. Identifying a kgroup entry as stable signifies that there are no current writes to any object in the kgroup and that the last leader-provided consistent_followers bitmap points to followers that have the last committed value for every object in the kgroup. Consequently, any of the consistent followers will return a value identical to the leader's value.

Second, after forwarding a read request to a follower (say, flwrA), the switch may receive a write request that modifies the object. The leader may replicate the write request to a majority of nodes that does not include flwrA. If the leader processes the write request before flwrA serves the read request, flwrA will return stale data. To avoid this case, the switch performs a safety check on every read reply coming from followers: it verifies that the kgroup is still stable, and that the sequence number in the read_reply is equal to the sequence number in the kgroup entry. If the sequence numbers do not match (which indicates that there are later writes to objects in the kgroup), the switch conservatively drops the read reply and forwards the request to the leader.

At all times, reads are linearizable in FLAIR.

6. Implementation

To demonstrate the benefits of the new approach, we prototyped FlairKV, a FLAIR-based key-value store built atop Raft [32]. We chose Raft due to its adoption in production systems [47, 48, 49, 50, 51], and the availability of standalone production-quality implementations [52].

6.1. Storage System Implementation

We have implemented FlairKV, including all switch data plane features, the FLAIR module, leaders' and followers' modifications, and the client library. We extended the Raft's follower code to implement an advance_then_read() function. We extended the leader to notify the *lflair* module as soon as it gets elected, and to extract its commit_index, the list of followers with a commit_index equal to the leader's commit_index, and the list of uncommitted writes. We extended the write reply with the list of followers which acknowledged the write. We implemented the leader lease optimization [10, 26] and modified Raft's client library to add the FLAIR header to client requests.

6.2. Switch Data Plane Implementation

The switch data plane is written in P4 v14 [33] and is compiled for Barefoot's Tofino ASIC [36], with Barefoot's P4Studio software suite [53]. Our P4 code defines 30 tables and 12 registers: six for the session array and six for the kgroup array. The kgroup array has 4K entries. Larger number of kgroups had negligible effect on performance. In total,

our implementation uses less than 5% of the on-chip memory available in the Tofino ASIC, leaving ample resources to support other switch functionalities or more FlairKV instances. The rest of this section discusses optimizations implemented in FlairKV to cope with the strict timing and memory constraints of P4 and switch ASIC.

Heartbeats implementation. The leader and the switch exchange periodic heartbeats. If the switch_stepdown period passes without receiving a leader heartbeat, the switch deactivates the session. Instead of running a process in the controller to continuously track heartbeats, the switch monitors missed heartbeats as part of the validation step in the processing pipeline. The switch keeps track of the timestamp of the last heartbeat received in the session array (Listing 1). When processing any FLAIR packet, the switch computes the difference between the current time and the last heartbeat timestamp; if the difference is larger than switch_stepdown, the switch deactivates the session, making the system unavailable until the leader starts a new session.

Forwarding logic translates the consistent followers' bitmap to follower IP addresses. Storing the IP addresses of consistent followers for every entry in the kgroup array significantly increases the memory footprint. Moreover, randomly selecting a follower from the list while avoiding inconsistent ones is tricky given the P4 and current ASIC challenges (Section 2.2). Instead, the FlairKV leader encodes the follower status in a one-byte consistent_followers bitmap (Listing 1). Replicas are ordered in a list. If the least significant bit in the consistent_follower bitmap is set, this indicates that the first replica in the list is consistent, and so forth.

When forwarding a read request, the switch translates the encoded bitmap of consistent followers to select one follower; Figure 6 shows the translation process. The consistent_followers bitmap is used as an index to the translation table. Each entry in the table has an action that randomly selects a number that is then used as an index to the IP addresses table.

This design has two benefits: it significantly reduces the memory footprint of the kgroup array, and it can be accelerated using P4 "action profiles" [54].

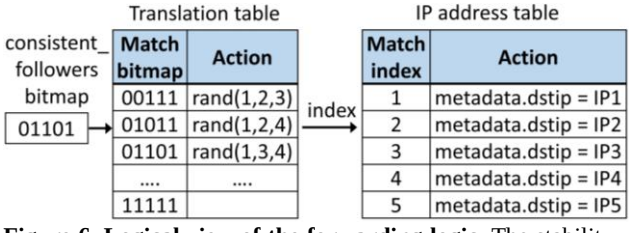


Figure 6. Logical view of the forwarding logic. The stability bitmap matches an entry in the translation table and executes the corresponding action, generating an index of the selected destination's IP address. Using the index, the IP address table sets the destination's IP address in the metadata.

Load balancing. In addition to the aforementioned random load-balancing technique (Figure 6), we implemented two load-aware techniques:

- *Leader avoidance*. Our benchmarking revealed that the write operation takes 35 times longer than a read operation; most of this overhead is borne by the leader. Consequently, this load-balancing technique avoids sending read requests to the leader for stable kgroups if there are any writes in the system. The aim is to reduce the leader load, as it is already busy serving writes and serving reads for unstable kgroups.
  To implement this technique, we extended the session array entry with a 64-bit largest_reply_seq_num field, which tracks the largest sequence number ever reported in a write_reply. If largest_reply_seq_num does not equal session_seq_num, then there are pending writes in the system and the leader should not be burdened with any reads to stable kgroups.
- *Follower load awareness*. This technique distributes the load across followers proportionally to their load in the last *n* seconds. This technique is especially useful for deployments that use heterogeneous hardware, experience workload variations, or deploy more than one replica (i.e., replicas for different ranges of keys) on the same machine. In our design, followers report the length of the request queue in every heartbeat. Every second, the leader calculates the average queue length for each follower and assigns proportional weights to each follower. The leader updates the translation table to reflect these weights. For instance, if follower 1 should receive double the load of any other replica, the action for a bitmap 00111 will be rand(1, 1, 2, 3), doubling the chance replica 1 is selected.

Register access logic. Each stage has its own dedicated registers, and a register can be accessed only once in a stage. This restriction complicates FlairKV's logic, as different packet types (e.g., read and write_reply) must access the same registers at different stages in the pipeline. To cope with this restriction, FlairKV adds a dedicated table to access each register. Figure 7 shows an example of an action table for accessing register *r1*. Our code aggregates the information about all possible modes of accessing *r1* in the packet's metadata, including the access type (read or write), the index, and which data should be written or where the value should be read to. We then use a dedicated match-action table (Figure 7) to perform the actual read or write operation to/from the register in a single stage with a single invocation of the table. This approach has the additional benefit of reducing the number of stages.

Processing concurrent requests. The switch processes packets sequentially in a pipeline. Each pipeline stage processes one packet at a time. The switch may have multiple pipelines, each serving a subset of switch ports. FLAIR uses

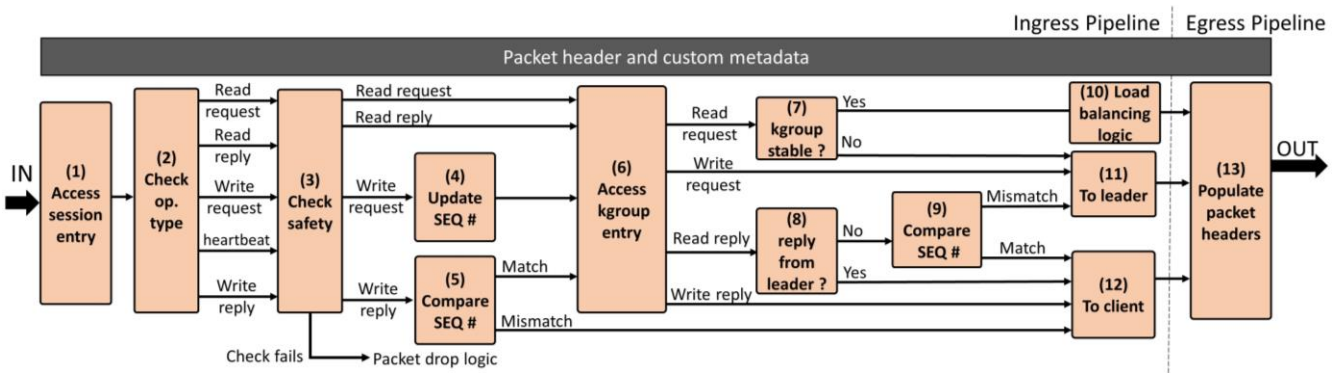


Figure 8. Logical view of the FlairKV switch data plane.

a single ingress pipeline and all egress pipelines. If a FLAIR packet is received on a different ingress pipeline, the packet is recirculated [54] to the FLAIR pipeline.

6.3. Putting the Switch Pipeline Together

Figure 8 shows the pipeline layout in the switch data plane and the flow for a FlairKV packet. The pipeline starts by reading the session information (1 in Figure 8) and adding it to the packet metadata. Then the it extracts the operation type (2) and validates the request (3) by verifying the session id. If the packet has an older session id the packet is dropped. Further, in the validation stage the switch confirms that it did not miss leader heartbeats in the last switch_stepdown period (Section 4.6), else it deactivates the session.

Read requests access the kgroup array (6), and if the group is stable, the request is forwarded to a load-balancing logic (10) that implements the forwarding logic (Section 6.2); otherwise, it is sent to the leader.

If a read reply is from the leader, it is forwarded to the client (12). If it is from a follower, the pipeline performs the safety check (9) and, if it suspects the reply is stale, drops the reply, then resubmits the read request to the leader (11).

Write requests update the session_seq_num (4) and the kgroup entry (6), then are sent to the leader (11).

Write replies compare the sequence number of the reply to the one in the kgroup entry (5); if they match, the kgroup entry is updated (6) and the pipeline forwards the reply to the client (12).

The egress pipeline (13) has one logical stage that populates the header fields (e.g., SEQ number, SID, etc.) using the data available in the packet's metadata.

```
If(type == read){
   mtdata.r1type = 0;
   mtdata.r1idx = idx;}
If (type == write){
   mtdata.r1type = 1;
   mtdata.r1idx = idx;
   mtdata. r1val = val;}
apply(r1_table)
```

Record access information in the packet metadata at different stages

access *r1* in a single stage

Match	Action
mtdata.r1type == 0	mtdata.r1val = reg1[mtdata.r1idx]
mtdata.r1type == 1	reg1[mtdata.r1idx] = mtdata.r1val

Register r1 access table (**r1_table**)

Figure 7. Register access table. P4 code aggregates access information that is used by a dedicated register access table.

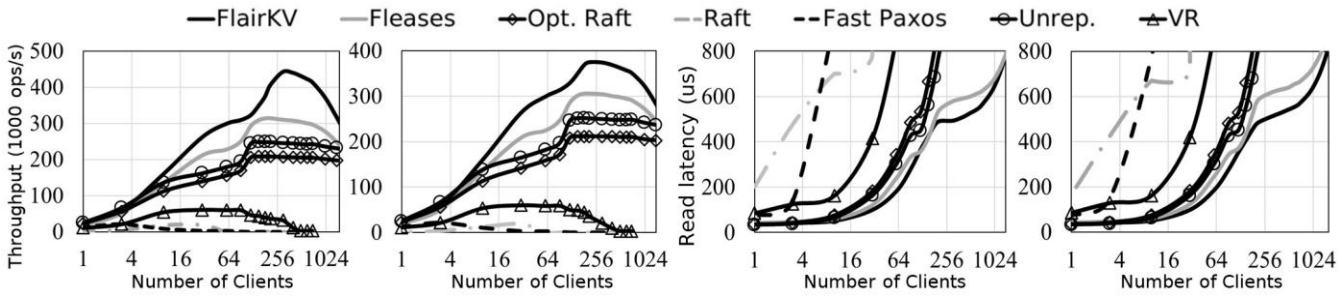
7. Evaluation

We compare our prototype with previous approaches in terms of throughput and latency (§7.1) with different workload skewness (§7.2) and read/write ratios (§7.3). In Appendix A, we evaluate FlairKV scalability, the benefits of different load-balancing techniques, and FlairKV's performance during failure recovery.

Testbed. We conducted our experiments using a 13-node cluster. Each node has an Intel Xeon Silver 10-core CPU, 48GB of RAM, and 100Gbps Mellanox NIC. The nodes are connected to an Edgecore Wedge 100 ×32BF switch with 32 100Gbps ports. The switch has Barefoot's Tofino ASIC, which is P4 programmable. Unless otherwise specified, three machines ran the server code, while the other 10 machines generated the workload.

Alternatives. We compare the throughput and latency of the following designs and optimizations:

- **Leader-based**. We used two leader-based protocol implementations: LogCabin, the original implementation of Raft (**Raft**), and an implementation of Viewstamped replication (**VR**) [28]. Raft and VR implement a batching optimization which batches and replicates multiple log entries in a single round trip.
- **Optimized Leader-based (Opt. Raft).** Our benchmarking revealed that the original Raft implementation could not utilize the resources of our cluster. We implemented two main optimizations: first, we changed the request-processing logic from an event-driven to a thread-pool design, as our benchmarking indicated a thread-pool performs better; second, we implemented the leader-lease optimization. These changes significantly improved Raft's performance.
- **Quorum-based reads (Fast Paxos).** An alternative to the leader-based design is the quorum design [43, 44, 55].



(a) Throughput - Uniform (b) Throughput -Zipf (c) Latency-Uniform (d) Latency-Zipf

Figure 9. Throuhgput and Latency while varying the number of clients. The figures show the throughput and the average latency for different number of clients for workload B for the uniform distribution (a, c), and for the Zipf distribution (b, d).

Typically, client read requests are sent to all followers, and each follower responds directly to the client. The client waits for a reply from a supermajority [55] before completing a read. We used a Fast Paxos implementation that implements only the normal case [28].

- **Follower-lease optimization (FLeases).** Similar to MegaStore [3], the leader grants read leases to all followers. Before serving a write, the leader revokes all leases, processes the write operation, and then grants a new lease to followers. The lease's grant/revoke messages are piggybacked on the consensus protocol messages. However, writes should be processed by *all* followers before replying to the client. In our experiments, if a follower receives a read request for an object for which it does not have an active lease, it forwards the request to the leader. MegaStore applications typically partition the keys into thousands of groups, each group contains logically-related keys [3] (e.g., a key group per blog [3]). We partitioned the keys into 4K groups (the same number of kgroups in FlairKV), and followers get a lease per group. Clients randomly select a follower for each read request and send the request directly to it.
- **Unreplicated/NOPaxos (Unrep.).** As a baseline, the unreplicated configuration deploys Optimized-Raft (discussed above) on a single node. The single node stores the data set and serves all operations without replication.

  This configuration also represents the best possible performance of the network-optimized NOPaxos [43] protocol. NOPaxos uses a network switch to order and multicast read and write operations to all replicas. An operation is successful if the majority accepts a write or returns the same value for a read. Consequently, NOPaxos read performance is limited by the slowest node in the majority of nodes. NOPaxos evaluation shows that the best throughput and latency the protocol can achieve are within 4% that of an unreplicated system [43].
- **FlairKV.** Unless otherwise specified, we used FlairKV with the leader-avoidance load-balancing technique.

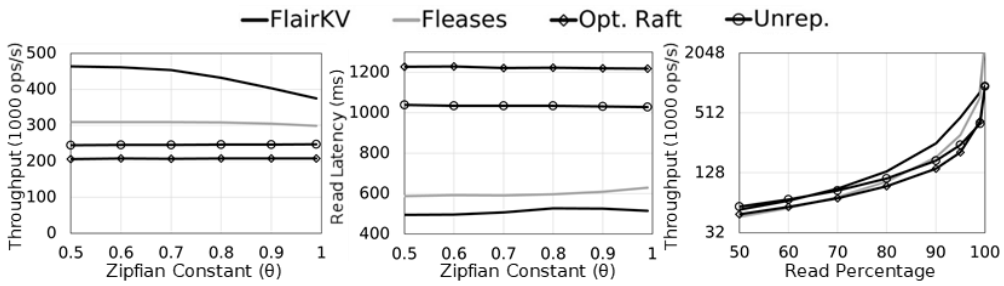
We benchmarked every system and selected a configuration that maximized its performance. We stored all data in memory. In all experiments, all systems' performance (with the exception of FastPaxos) was stable with a standard deviation less than 1%.

Workload. We used synthetic benchmarks and the YCSB benchmark [56] to evaluate the performance of all systems. In our evaluation, we considered both uniform and skewed workloads. The skewed workload follows the Zipf distribution with a skewness parameter of 0.99. We also used the YCSB benchmark. We experimented with 100,000 and 1 million keys. We present the results with 100,000 keys as, in skewed workloads, the fewer number of hot keys increased the chance of having concurrent requests accessing the same key (i.e. is less favorable for FlairKV). FlairKV bring slightly higher performance benefit when using 1 million keys than 100,000 keys. The key size is 24 bytes and the hash of the key string is used as the key in the FLAIR protocol. The value size is 1KB.

7.1. Performance Evaluation

We compared the seven systems using YCSB workload B (95:5 read:write ratio) while varying the number of clients, with uniform and skewed workload distribution. Figure 9 shows the throughput and average latency with a uniform and skewed distributions. With the uniform distribution (Figure 9 (a) and (c)), FlairKV achieves up to 42% higher throughput and 23.7% lower average latency than FLeases, and 1.3 to 2.3 times higher throughput and 1.5 to 2.4 times lower latency compared to optimized Raft and unreplicated setup. Fast Paxos, Raft, and VR, achieve the lowest throughput and highest latency as these systems contact the majority of nodes for every read.

FlairKV achieved better performance than FLeases for three reasons. First, FlairKV uses the leader-avoidance load-balancing technique, which reduces the load on the leader when there are writes, thereby accelerating writes and shortening the time period in which kgroups are marked unstable. This approach is effective as writes take almost 35 times longer than reads in Opt.Raft, and 30 times longer in the unreplicated setup. We recorded the number of read requests served by the leader. For instance, with 300 clients (Figure 9.a) the leader served 2% of the reads in FlairKV (those are reads to unstable kgroups), while it served 34% of



(a) Throughput (b) Read latency

Figure 10. Throughput and Latency while varying skewness. The figures show the throughput (a) and the average latency (b) for different zifpian constants for a uniform workload B with 300 clients.

Figure 11. Throughput while varying the read ratio. Using uniform workload B.

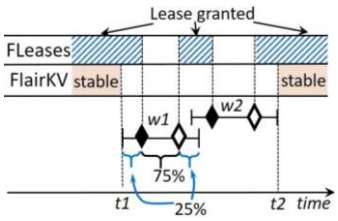


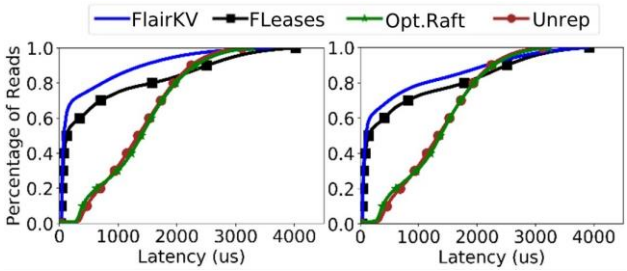
Figure 12. Subtle effect of FLAIR. FLeases may grant leases for up to 25% more time compared to FlairKV. Bars mark the time from the moment a switch receives a write request (w1 or w2) until it receives a corresponding reply.**12**

the reads in FLeases. We note that the leader-avoidance technique cannot be applied to FLeases which tasks the clients with selecting a follower to send the read request to. This technique requires accurate information about the current load of the leader and which followers are stable which are not available to clients.

Second, in FLeases, when an object is not stable, if a client sends a request to a follower, the follower will redirect the request to the leader, increasing overhead and incurring extra latency. Unlike FLeases, FlairKV switch knows if an object is not stable and forwards read requests for that object directly to the leader. The third reason which had a minor impact when using 3 replicas is that the write operation in FLeases need to reach all followers, while FlairKV writes only need a majority.

Optimized-Raft's performance is better than that of Raft, VR, and FastPaxos. The unreplicated deployment slightly improves throughput and latency over Optimized-Raft by avoiding the replication overhead for write operations. These two systems still lag behind FlairKV as they only utilize a single node (the leader) for serving all reads and writes.

Figure 9 (b) and (d) show the throughput and average latency with a skewed workload (Zifpian constant of 0.99). The skewed workload results in higher contention and an increased frequency at which a read request finds a kgroup unstable for the popular keys. This contention reduces the chances of reading from followers. FlairKV leader served 21% of reads of which 1% are redirected from followers, while FLeases leader served 37% of reads. Even under the skewed workload, FlairKV still achieves the highest performance, up to 26% higher throughput and 18.1% lower latency than FLeases, and 1.5 to 1.8 times higher throughput and 2 to 2.4 times lower latency than optimized Raft and the unreplicated setup.



(a) B-Uniform (b) B-Zipf

Figure 13. Latency CDF. The figures show the latency CDF for reads under workload B using 300 clients with a uniform distribution (a), and a Zipf distribution with skewness of 0.99 (b). The lines for Opt. Raft and Unrep. almost overlap.

Latency evaluation. Figure 13.a shows the latency CDF of FlairKV, FLeases, OptRaft, and Raft. Under the uniform workload B with 300 clients (other workloads had similar results). FlairKV lowered the latency for the slowest 40% requests by at least 38% relative to FLeases. Under the Zipf workload (Figure 13.b), FlairKV lowered the slowest 50% of request by up to 35% relative to FLeases.

FLeases has higher latency as it incurs extra delay due to the load imbalance between nodes (e.g., the leader serves 41% of requests for workload B with Zipf distribution) and due to followers redirecting 4% of requests to the leader.

Under all workloads, FlairKV significantly improved operation's latency relative to OptRaft and Raft. The median latency of FlairKV is 2% of Raft's latency and 2-8% of OptRaft's latency.

7.2. Workload Skewness

We measured the impact of the workload skewness on throughput (Figure 10.a) and average latency (Figure 10.b) by varying the Zipfian constant from 0.5 to 0.99. FlairKV consistently achieves better performance: 1.26 to 2.25 times higher throughput and 1.13 to 2.48 times lower average latency compared to all other systems. We notice that as the skewness increases FlairKV and FLeases performance decreases as higher skewness increases contention on the few popular kgroups, making them unstable for longer time, and increases the number of requests the leaders have to process. Other systems performance is not noticeably affected by skewness.

We noticed high workload skewness affects FlairKV's performance more than FLeases. This is due to a subtle side

effect of FlairKV. When there are concurrent writes to the same kgroup, FlairKV will mark a group unstable from the moment the first request is processed by the switch until the last request to the kgroup is replied to ([*t1*, *t2*] in Figure 12). In FLeases, the lease revocation is piggybacked on the write replication step (black diamonds in Figure 12). Once the leader commits a write, it sends a commit notification and grants a new lease to the followers (white diamonds). Hence, FLeases may grant a lease between concurrent writes, creating more opportunity for serving reads from followers.

To further understand this effect, we tracked leases and the stability of kgroups under the skewed (factor of 0.99) write heavy YCSB workload A (1:1 read:write ratio). We noticed that while 29% of reads found the kgroup unstable in FlairKV, only 4% of reads in FLeases reached a follower that did not have a lease. We further profiled the write operation path and found that FLeases revokes leases for 75% of the write operation time (Figure 12), 25% shorter than the period FlairKV marks a kgroup unstable. Despite this subtle effect FlairKV leader still has lighter load, it served 29% of reads compared to 37% served by the FLeases leader. Notwithstanding this effect FlairKV still brings 17% to 26% performance improvement even under skewed workloads.

7.3. Read/Write Ratio

Figure 11 shows the effect of the ratio of reads to writes on systems' performance with a uniform workload B. Compared to FLeases, FlairKV has up to 1.5 times higher throughput for all read to write ratios, with the exception of the read-only workload in which their performance is comparable. FlairKV has 1.25 to 2.8 times higher throughput compared to the Opt. Raft. Compared to the unreplicated setup, FlairKV has up to 2.8 times higher throughput for workloads with 70% reads or more and a comparable performance under write heavy workloads (read ratio 50-70%).

7.4. Fault Tolerance

To demonstrate FlairKV fault tolerance techniques, we measured the system throughput using workload C under three failure scenarios: switch, leader, and follower failure. Our evaluation shows that FlairKV can efficiently recover from these failures. On a switch failure the system is not available for 750ms. This period is dominated by waiting for three heartbeats (250ms each). The actual recovery protocol takes 2RTT and transfers only 4KB of data. Appendix A shows the detailed fault tolerance evaluation.

8. Related Work

Network-accelerated systems. Recent projects have utilized SDN capabilities to provide load balancing [57, 58, 59], access control [60], seamless virtual machine migration [61], and improving system security, virtualization, and network efficiency [62]. SwitchKV [31] uses SDN capabilities to route client requests to the caching node serving the key. A central controller populates the forwarding rules to invalidate routes for objects that are being modified and installs routes for newly cached objects. NetCache [30] proposes using the limited switch memory as a look-through cache for key-value stores.

Network-accelerated consensus. A number of recent efforts leverage SDN's capabilities to optimize consensus protocols. Speculative Paxos [44] builds a mostly ordered multicast primitive and uses it to optimize the multi-Paxos consensus protocol. Network-ordered Paxos (NOPaxos) [43] leverages modern network capabilities to order multicast messages and add a unique sequence number to every client request. NOPaxos uses these sequence number to serialize operations and to detect packet loss. Speculative Paxos and NOPaxos are optimized for operations that update the log but not for read operations. NetChain [63] and NetPaxos [64] implement replication protocols on a group of switches. These protocols are suitable for systems that store only a few megabytes of data (e.g., 8MB in the NetChain prototype). Unlike FLAIR, these previous efforts do not optimize for read operations. Reads are still served by the leader or a quorum of replicas.

Consensus protocols optimized for the WAN. A number of consensus protocols are optimized for WAN deployments. Quorum leases [16] proposes giving a read lease to some of the followers; Unlike Megastore leases, when an object is modified, only the followers that have the lease are contacted. Quorum leases has a better performance than Megastore leases in WAN setups, but do not bring benefits when deployed in a single cluster [16]. Mencius [65] is a multi-leader protocol in which each leader controls part of the log. EPaxos [66] is a leaderless protocol where clients can submit request to any replica. Non-conflicting write can commit in one round trip, while conflicting writes will be resolved using Paxos.

CURP [67] optimizes the write operation through exploiting commutativity between concurrent writes. In data center deployments, CURP reads are served by the leader and hence are limited to a single node performance, in WAN deployment CURP applies a technique similar to FLeases.

9. Conclusion

We present FLAIR, a novel protocol that leverages the capabilities of the new generation of programmable switches to accelerate read operations without affecting writes or using leases. FLAIR identifies, at line rate, which replicas can serve a read request consistently, and implements a set of load-balancing techniques to distribute the load across consistent replicas. We detailed our experience building FlairKV and presented a number of techniques to cope with the restrictions of the current programmable switches. We hope our experience informs a new generation of distributed

systems that co-design network protocols and functionalities with system operations.

References

[1] H. Takruri, I. Kettaneh, A. Alquraan et al., "FLAIR: Accelerating Reads with Consistency-Aware Network Routing," in 17th USENIX Symposium on Networked Systems Design and Implementation, 2020.

[2] Ibrahim Kettaneh, Ahmed Alquraan, Hatem Takruri, Ali José Mashtizadeh and Samer Al-Kiswany, "Accelerating Reads with In-Network Consistency-Aware Load Balancing," IEEE Transaction on Networking, 2021.

[3] J. Baker, C. Bond, J. C. Corbett et al., "Megastore: Providing scalable, highly available storage for interactive services," in Proceedings of the Conference on Innovative Data system Research (CIDR), 2011.

[4] P. Hunt, M. Konar, F. P. Junqueira et al., "ZooKeeper: wait-free coordination for internet-scale systems," in Proceedings of the USENIX annual technical conference, Boston, MA, 2010.

[5] B. Calder, J. Wang, A. Ogus et al., "Windows Azure Storage: a highly available cloud storage service with strong consistency," in Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles (SOSP), Cascais, Portugal, 2011, doi: 10.1145/2043556.2043571.

[6] J. C. Corbett, J. Dean, M. Epstein et al., "Spanner: Google's globally-distributed database," in Proceedings of the USENIX conference on Operating Systems Design and Implementation (OSDI), Hollywood, CA, USA, 2012: USENIX Association.

[7] N. Bronson, Z. Amsden, G. Cabrera et al., "TAO: Facebook's distributed data store for the social graph," in Proceedings of the USENIX Technical Conference, San Jose, CA, 2013: USENIX Association.

[8] H. Attiya and J. Welch, Distributed Computing: Fundamentals, Simulations and Advanced Topics. John Wiley & Sons, Inc., 2004.

[9] L. Lamport, "Paxos made simple," ACM Sigact News, vol. 32, no. 4, pp. 18-25, 2001.

[10] D. Ongaro and J. Ousterhout, "In search of an understandable consensus algorithm," in Proceedings of the USENIX Annual Technical Conference, Philadelphia, PA, 2014: USENIX Association.

[11] F. P. Junqueira, B. C. Reed, and M. Serafini, "Zab: High-performance broadcast for primary-backup systems," in Proceedings of IEEE/IFIP International Conference on Dependable Systems&Networks, 2011: IEEE Computer Society, doi: 10.1109/dsn.2011.5958223.

[12] B. Liskov and J. Cowling, "Viewstamped replication revisited," Technical Report MIT-CSAIL-TR-2012-021, MIT, 2012.

[13] J. Shute, R. Vingralek, B. Samwel et al., "F1: a distributed SQL database that scales," Proc. VLDB Endow., vol. 6, no. 11, pp. 1068-1079, 2013, doi: 10.14778/2536222.2536232.

[14] B. Atikoglu, Y. Xu, E. Frachtenberg et al., "Workload analysis of a large-scale key-value store," presented at the Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems, London, England, UK, 2012.

[15] C. Gray and D. Cheriton, "Leases: an efficient fault-tolerant mechanism for distributed file cache consistency," in Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), 1989: ACM, doi: 10.1145/74850.74870.

[16] I. Moraru, D. G. Andersen, and M. Kaminsky, "Paxos Quorum Leases: Fast Reads Without Sacrificing Writes," in Proceedings of the ACM Symposium on Cloud Computing, Seattle, WA, USA, 2014: ACM, doi: 10.1145/2670979.2671001.

[17] J. Terrace and M. J. Freedman, "Object storage on CRAQ: high-throughput chain replication for read-mostly workloads," presented at the Proceedings of the 2009 conference on USENIX Annual technical conference, San Diego, California, 2009.

[18] M. Burrows, "The Chubby lock service for loosely-coupled distributed systems," in Proceedings of the Symposium on Operating Systems Design and Implementation (OSDI), Seattle, Washington, 2006: USENIX Association.

[19] "Swift's documentation." https://docs.openstack.org/swift/stein/index.html (accessed April 14, 2019.

[20] "Redis." https://redis.io (accessed April 14, 2019.

[21] "Apache Cassandra." https://cassandra.apache.org (accessed April 14, 2019.

[22] G. DeCandia, D. Hastorun, M. Jampani et al., "Dynamo: amazon's highly available key-value store," in Proceedings of the ACM SIGOPS Symposium on Operating Systems Principles (SOSP), Stevenson, Washington, USA, 2007: ACM, doi: 10.1145/1294261.1294281.

[23] D. B. Terry, V. Prabhakaran, R. Kotla et al., "Consistency-based service level agreements for cloud storage," in Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), Farminton, Pennsylvania, 2013: ACM, doi: 10.1145/2517349.2522731.

[24] B. F. Cooper, R. Ramakrishnan, U. Srivastava et al., "PNUTS: Yahoo!'s hosted data serving platform,"

Proc. VLDB Endow., vol. 1, no. 2, pp. 1277-1288, 2008, doi: 10.14778/1454159.1454167.
[25] M. Poke and T. Hoefler, "DARE: High-Performance State Machine Replication on RDMA Networks," in Proceedings of the International Symposium on High-Performance Parallel and Distributed Computing, Portland, Oregon, USA, 2015, 2749267: ACM, pp. 107-118, doi: 10.1145/2749246.2749267.
[26] T. D. Chandra, R. Griesemer, and J. Redstone, "Paxos made live: an engineering perspective," in Proceedings of the annual ACM symposium on Principles of distributed computing, Portland, Oregon, USA, 2007: ACM, doi: 10.1145/1281100.1281103.
[27] D. Mazieres, "Paxos made practical," Technical Report on http://www.scs.stanford.edu/~dm/home/papers/paxos.pdf, 2007.
[28] "NOPaxos consensus protocol." https://github.com/UWSysLab/NOPaxos (accessed April 14, 2019.
[29] S. Al-Kiswany, S. Yang, A. C. Arpaci-Dusseau et al., "NICE: Network-Integrated Cluster-Efficient Storage," in Proceedings of the International Symposium on High-Performance Parallel and Distributed Computing, Washington, DC, USA, 2017: ACM, doi: 10.1145/3078597.3078612.
[30] X. Jin, X. Li, H. Zhang et al., "NetCache: Balancing Key-Value Stores with Fast In-Network Caching," in Proceedings of the Symposium on Operating Systems Principles (SOSP), Shanghai, China, 2017: ACM, doi: 10.1145/3132747.3132764.
[31] X. Li, R. Sethi, M. Kaminsky et al., "Be fast, cheap and in control with SwitchKV," in Proceedings of the Usenix Conference on Networked Systems Design and Implementation (NSDI), Santa Clara, CA, 2016: USENIX Association.
[32] "LogCabin storage system." https://logcabin.github.io (accessed April 14, 2019.
[33] "P4." https://p4.org (accessed April 14, 2019.
[34] D. Mazieres, "Paxos made practical," Technical Report on http://www.scs.stanford.edu/~dm/home/papers/paxos.pdf.
[35] L. Lamport, "The part-time parliament," ACM Trans. Comput. Syst., vol. 16, no. 2, pp. 133-169, 1998, doi: 10.1145/279227.279229.
[36] "Barefoot Tofino." https://www.barefootnetworks.com/products/brief-tofino/ (accessed April 14, 2019.
[37] "Cavium / XPliant." https://origin-www.marvell.com/documents/netpxrx94dcdhk8sksbp/ (accessed April 14, 2019.
[38] "High-Capacity StrataXGS® Trident 3 Ethernet Switch Series." https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56870-series (accessed September 9, 2019.
[39] P. Bosshart, D. Daly, G. Gibb et al., "P4: programming protocol-independent packet processors," SIGCOMM Comput. Commun. Rev., vol. 44, no. 3, pp. 87-95, 2014, doi: 10.1145/2656877.2656890.
[40] "Data Center: Load Balancing Data Center." https://learningnetwork.cisco.com/docs/DOC-3438 (accessed April 14, 2019.
[41] L. A. Barroso and U. Hoelzle, The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. Morgan and Claypool Publishers, 2009, p. 120.
[42] N. McKeown, T. Anderson, H. Balakrishnan et al., "OpenFlow: enabling innovation in campus networks," SIGCOMM Comput. Commun. Rev., vol. 38, no. 2, pp. 69-74, 2008, doi: 10.1145/1355734.1355746.
[43] J. Li, E. Michael, N. K. Sharma et al., "Just say no to paxos overhead: replacing consensus with network ordering," in Proceedings of the USENIX conference on Operating Systems Design and Implementation (OSDI), Savannah, GA, USA, 2016: USENIX Association.
[44] D. R. K. Ports, J. Li, V. Liu et al., "Designing distributed systems using approximate synchrony in data center networks," in Proceedings of the USENIX Conference on Networked Systems Design and Implementation (NSDI), Oakland, CA, 2015: USENIX Association.
[45] "TLA+ Language." https://lamport.azurewebsites.net/tla/tla.html (accessed April 14, 2019.
[46] D. Ongaro. "Raft TLA+ Specification." https://github.com/ongardie/raft.tla (accessed 2019).
[47] "etcd: Distributed reliable key-value store for the most critical data of a distributed system." https://github.com/etcd-io/etcd (accessed April 14, 2019.
[48] "RethinkDB: the open-source database for the realtime web." https://www.rethinkdb.com/ (accessed April 14, 2019.
[49] "Open Network Operating System (ONOS) - Cluster Coordination." https://wiki.onosproject.org/display/ONOS/Cluster+Coordination (accessed.
[50] "Apache Kudu - Fast Analytics on Fast Data." https://kudu.apache.org/ (accessed April 14, 2019.
[51] "Hashicorp Raft implementation." https://github.com/hashicorp/raft (accessed April 14, 2019.
[52] "The Raft Consensus Algorithm." https://raft.github.io/ (accessed April 14, 2019.

[53] "Barefoot P4 Studio." https://www.barefootnetworks.com/products/brief-p4-studio/ (accessed April 14, 2019.
[54] "P4 v16 Portable Switch Architecture (PSA)." https://p4.org/p4-spec/docs/PSA-v1.0.0.html (accessed April 14, 2019.
[55] L. Lamport, "Fast paxos," Distributed Computing, vol. 19, no. 2, pp. 79-103, 2006.
[56] "Yahoo! Cloud Serving Benchmark in C++, a C++ version of YCSB." https://github.com/basicthinker/YCSB-C (accessed April 14, 2019.
[57] B. Cully, J. Wires, D. Meyer et al., "Strata: High-performance scalable storage on virtualized non-volatile memory," in Proceedings of the USENIX Conference on File and Storage Technologies (FAST), 2014, pp. 17-31.
[58] N. Handigol, M. Flajslik, S. Seetharaman et al., "Aster* x: Load-balancing as a network primitive," in GENI Engineering Conference (Plenary), 2010, pp. 1-2.
[59] R. Wang, D. Butnariu, and J. Rexford, "OpenFlow-based server load balancing gone wild," in Proceedings of the USENIX conference on Hot topics in management of internet, cloud, and enterprise networks and services, Boston, MA, 2011: USENIX Association.
[60] A. K. Nayak, A. Reimers, N. Feamster et al., "Resonance: dynamic access control for enterprise networks," in Proceedings of the ACM workshop on Research on enterprise networking, Barcelona, Spain, 2009: ACM, doi: 10.1145/1592681.1592684.
[61] A. J. Mashtizadeh, M. Cai, G. Tarasuk-Levin et al., "XvMotion: unified virtual machine migration over long distance," in Proceedings of the USENIX Annual Technical Conference, Philadelphia, PA, 2014: USENIX Association.
[62] A. Lara, A. Kolasani, and B. Ramamurthy, "Network innovation using openflow: A survey," IEEE communications surveys & tutorials, vol. 16, no. 1, pp. 493-512, 2014.
[63] X. Jin, X. Li, H. Zhang et al., "Netchain: scale-free sub-RTT coordination," in Proceedings of the USENIX Conference on Networked Systems Design and Implementation (NSDI), Renton, WA, USA, 2018: USENIX Association.
[64] H. T. Dang, D. Sciascia, M. Canini et al., "NetPaxos: consensus at network speed," in Proceedings of the ACM SIGCOMM Symposium on Software Defined Networking Research, Santa Clara, California, 2015: ACM, doi: 10.1145/2774993.2774999.
[65] Y. Mao, F. P. Junqueira, and K. Marzullo, "Mencius: building efficient replicated state machines for WANs," presented at the Proceedings of the 8th USENIX conference on Operating systems design and implementation, San Diego, California, 2008.
[66] I. Moraru, D. G. Andersen, and M. Kaminsky, "There is more consensus in Egalitarian parliaments," presented at the Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, Farminton, Pennsylvania, 2013.
[67] S. J. Park and J. Ousterhout, "Exploiting commutativity for practical fast replication," in 16th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 19), 2019, pp. 47-64.

A. Additional Evaluation

A.1 Load-balancing Performance Evaluation

We measured the system throughput using the following three configurations of FlairKV (detailed in Section 6.2)

- *FlairKV-Rand* selects a follower or the leader at random. Consequently, read requests for stable kgroups are uniformly spread across the followers and the leader.
- *FlairKV-LA* applies the leader-avoidance technique.
- *FlairKV-LA+FL* uses both leader-avoidance and follower load-awareness techniques.

Figure 14.a shows the performance improvement produced by the leader-avoidance technique. In this experiment, we used workload B with uniform key popularity distribution. The results show that FlairKV-LA throughput is higher by 40% than FlairKV-Rand throughput, as it accelerates writes and reduces the period in which kgroups are marked unstable. FlairKV-LA+FL had comparable performance to FlairKV-LA as nodes are homogenous.

Figure 14.b evaluates the benefits of using the follower-load awareness technique (Section 6.2). This technique helps in deployments with heterogeneous hardware and with load variance. To emulate such scenarios, we manually reduced the CPU frequency for one follower by 10%. We used the read-only workload C with uniform distribution to avoid write operations (as those give advantage to the leader-avoidance technique). FlairKV-LA+FL had 17% higher throughput relative to the other configurations, as it distributes the load proportionally to the node's request queue length. Furthermore, we noticed that FlairKV-LA+FL reduces latency by 10%. FlairKV-LA and FlairKV-Rand are equivalent under the read-only workload.

A.2 Scalability

To demonstrate FlairKV scalability, we measured the system throughput using a read-only YSCB workload C while varying the number of replicas (Figure 15). The figure shows that FlairKV throughput scales linearly with the number of replicas, reaching 5.4 million request per second with 6 followers. We notice that the system achieves much higher performance under the read-only workload mainly due to the lower operation overhead (as writes take 30 times longer even without accounting for the replication overhead).

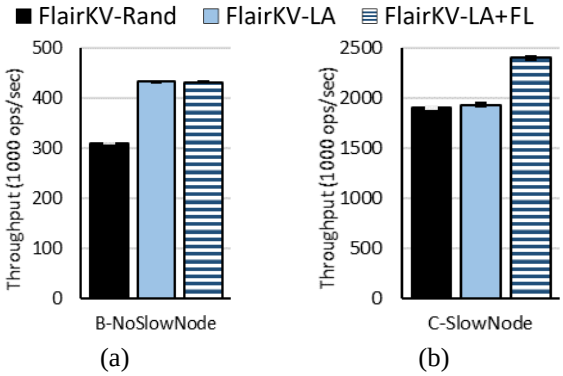


Figure 14. Throughput using different load-balancing techniques. (a) Uses workload B without slowing any follower and (b) uses workload C and slows one of the followers.

A.3 Fault Tolerance

To demonstrate FlairKV fault tolerance techniques, we measured the system throughput using workload C under three failure scenarios: switch, leader, and follower failure.

Switch Failure. We ran FlairKV at peak throughput for 35 seconds (Figure 16). At the 10s mark, the controller emulated a switch failure by wiping out the switch registers and installing rules to drop switch heartbeats. After missing 3 heartbeats, the leader suspects that the switch has failed and starts a new session. During this process, the switch is inactive, which causes the throughput to drop to zero for 750ms. Afterwards, the switch resumes normal operations.

Leader Failure. Figure 17 shows FlairKV throughput during the leader failure. We ran FlairKV at peak throughput for 35 seconds. At the 10s mark, we kill the leader process. Write requests fail, but the switch continues to forward read requests to followers. After missing 3 heartbeats the switch deactivates the session, and the throughput drops to zero. After 6 heartbeats, the followers elect a new leader that starts a new session. The system resumes its operation with one leader and one follower.

Follower Failure. We ran FlairKV at peak throughput for 35 seconds (Figure 18). At the 10s mark, we kill a follower process. This causes a drop in throughput as fewer replicas are available to serve read requests. The switch keeps forwarding client requests to the failed follower until the leader updates the switch. The dip in throughput at the second 10 is because we use closed-loop clients and some of the clients block waiting for the failed replica before timing out and retrying. Afterwards, the system throughput drops by 33% due to the loss of one follower.

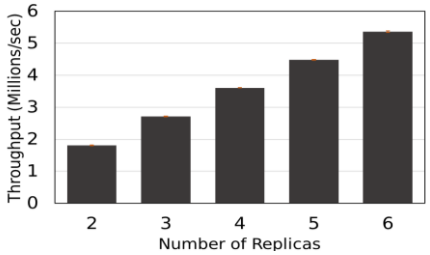


Figure 15. FlairKV scalability with different number of replicas

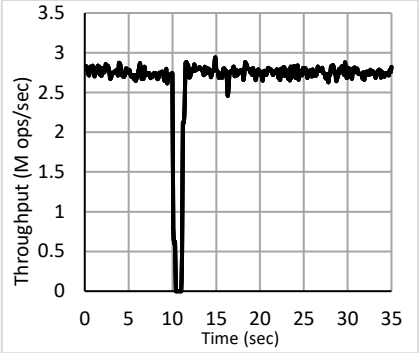


Figure 16. FlairKV throughput during a switch failover.

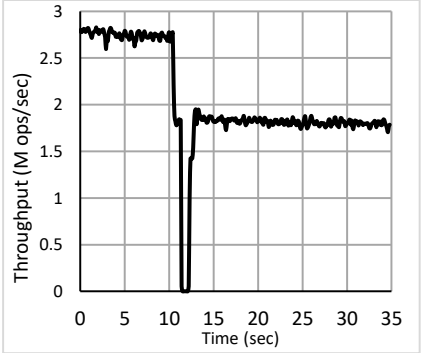


Figure 17. FlairKV throughput during leader failover.

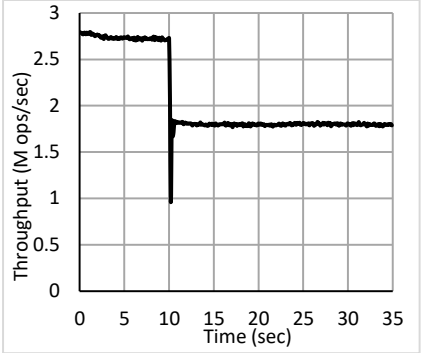


Figure 18. FlairKV throughput during a follower failure.

B. TLA+ Formal Specification

MODULE *FLAIR*

Formal specification for *FLAIR*.

EXTENDS *Naturals*, *FiniteSets*, *Sequences*, *TLC*, *Reals*

Constants

The set of replicas
CONSTANTS *Replicas*

Replicas states.
CONSTANTS *Follower*, *Leader*

Replicas running state
CONSTANTS *ReplicaUpState*, *ReplicaDownState*

CONSTANTS *SwitchIp*, Ip address of the switch
SwitchKGroupsNum, Number of key groups
SwitchStateActive,
SwitchStateInactive

Message Types
CONSTANTS *ClientReadRequest*, Read request type
ClientWriteRequest, Write request type
WriteResponse, Write response type
ReadResponse, read response type
Read request from the switch to a replica
InternalReadRequest,
Write request from the switch to a replica
InternalWriteRequest,
A request from the leader to the replicas
to append a *log* entry
AppendEntriesRequest,
An append entry resposne from a replica
to the leader
AppendEntriesResponse

The set of possible keys and values in a client request
CONSTANTS *ValueSpace*, *KeySpace*

Special reserved value
CONSTANTS *Nil*

Variables

Replica *vars*

VARIABLE *state*, The replica's state (Follower, or *Leader*).
log, Log of all committed and uncommitted operations
commitIndex, The index of the highest committed index in the *log*
currentTerm, Current term number
isActive, Is the replica up or down
replicaSession The latest switch id seen by the leader

$replicaVars \triangleq \langle state, log, commitIndex, currentTerm, isActive, replicaSession \rangle$

Leader *vars*. The following variables are used only on leaders:

VARIABLE *nextIndex*, The next entry to send to each follower.
matchIndex, The entry index for which a follower *log* matches the leader *log*. This used to calculate commit index
replicaKGroups A map from a key hash to the last received sequence number for the associated *KGroup*

$leaderVars \triangleq \langle nextIndex, matchIndex, replicaKGroups \rangle$

Switch *vars*

VARIABLES *switchKGourpArray*, Array to maintain information about each *KGroup*
switchSeqNum, The last sequence number assigned by the switch
switchTermId, The latest term number seen by the switch
switchLeaderId, Current leader id as seen by the switch
session, Current switch id nique value for each session
switchState Is the switch up or down

$switchVars \triangleq \langle switchKGourpArray, switchState, switchTermId, switchLeaderId, switchSeqNum, session \rangle$

Messages variables

VARIABLE *messages*, messages among replicas
msgsClientSwitch, messages from the client to the switch
msgsReplicasSwitch messages from between replicas and the switch

$msgsVars \triangleq \langle messages, msgsClientSwitch, msgsReplicasSwitch \rangle$

Proof *vars*, don't appear in implementations

VARIABLES *responsesToClient*

Set of all variables

$vars \triangleq \langle msgsVars,\ replicaVars,\ leaderVars,\ switchVars,\ responsesToClient \rangle$

Helpers

Helper to *Append* an element to a set

$AddToSet(set,\ element) \triangleq set \cup \{element\}$

Helper to remove an element to a set

$RemoveFromSet(set,\ element) \triangleq set \setminus \{element\}$

Helper to return the minimum value from a set,
or undefined if the set is empty.

$Min(s) \triangleq \text{CHOOSE } x \in s : \forall\, y \in s : x \leq y$

Helper to return the maximum value from a set,
or undefined if the set is empty.

$Max(s) \triangleq \text{CHOOSE } x \in s : \forall\, y \in s : x \geq y$

Helper to choose a random element from a set

$ChooseRandomly(set) \triangleq \text{CHOOSE } i \in set : \text{TRUE}$

Helper to return the index of the *KGroup* given a hash

$getIndexFromHash(hash) \triangleq (hash\%SwitchKGroupsNum) + 1$

The set of all quorums. This just calculates simple majorities, but the only
important property is that every quorum overlaps with every other.

$Quorum \triangleq \{i \in \text{SUBSET } (Replicas) : Cardinality(i) * 2 > Cardinality(Replicas)\}$

Helper to return the term of the last entry in a *log*,
or 0 if the *log* is empty.

$LastTerm(xlog) \triangleq \text{IF } Len(xlog) = 0 \text{ THEN } 0 \text{ ELSE } xlog[Len(xlog)].term$

Helper to find list of replicas that agree on a *log* index

$$AgreeIndex(index,\ logs,\ leaderId) \triangleq \\ \{leaderId\} \cup \{k \in Replicas : \\ \qquad \land\ k \neq leaderId \\ \qquad \land\ Len(logs[k]) \geq index \\ \qquad \land\ logs[k][index] = logs[leaderId][index]\}$$

Helper to return all *log* entries for specific key

$entriesForKey(s,\ key) \triangleq \{j \in \text{DOMAIN } log[s] : \land\ log[s][j].key = key\}$

return the max value of a set or *Nil* if the set is empty

$$indexOfLastEntry(entries) \triangleq \text{IF } Cardinality(entries) > 0 \\ \text{THEN } Max(entries) \text{ ELSE } Nil$$

helper function to get the *ID* of the leader
if there is a leader that is up, this function returns its *ID*
if there is no currect leader, this function finds the next leader based
on raft criteria and returns its *ID*

$$
\begin{array}{l}
getLeaderId \triangleq \\
\quad \text{LET } leadersList \triangleq \{s \in \text{DOMAIN } state : state[s] = \\
\qquad\qquad\qquad\qquad Leader \wedge isActive[s] = ReplicaUpState\} \\
\qquad runningReplicas \triangleq \{s \in Replicas : isActive[s] = ReplicaUpState\} \\
\qquad replicasWithNonEmptyLog \triangleq \{s \in runningReplicas : Len(log[s]) > 0\} \\
\qquad latestTerm \triangleq Max(\{log[s][Len(log[s])].term : \\
\qquad\qquad\qquad\qquad\qquad\qquad s \in replicasWithNonEmptyLog\}) \\
\qquad replicasWithLatestTerm \triangleq \{s \in replicasWithNonEmptyLog : \\
\qquad\qquad\qquad\qquad\qquad log[s][Len(log[s])].term = latestTerm\} \\
\qquad lengthOfLargestLog \triangleq Max(\{Len(log[s]) : s \in replicasWithLatestTerm\}) \\
\qquad replicasWithLongestLog \triangleq \{s \in replicasWithLatestTerm : \\
\qquad\qquad\qquad\qquad\qquad Len(log[s]) = lengthOfLargestLog\} \\
\qquad newLeaderId \triangleq \text{IF } Cardinality(replicasWithLongestLog) \geq 1 \\
\qquad\qquad\qquad\qquad \text{THEN CHOOSE } i \in replicasWithLongestLog : \text{TRUE} \\
\qquad\qquad\qquad\qquad \text{ELSE IF } Cardinality(runningReplicas) > 0 \\
\qquad\qquad\qquad\qquad \text{THEN CHOOSE } i \in runningReplicas : \text{TRUE} \\
\qquad\qquad\qquad\qquad \text{ELSE } Nil \\
\quad \text{IN IF } Cardinality(leadersList) > 0 \\
\qquad \text{THEN CHOOSE } s \in leadersList : \text{TRUE} \\
\qquad \text{ELSE } newLeaderId
\end{array}
$$

Helper to add a message to a set of messages.

$$Send(m) \triangleq messages' = AddToSet(messages, m)$$

Helper to remove a message from a set of messages.
Used when a replica is done processing a *replicaVarsmessage*.

$$Discard(m) \triangleq messages' = RemoveFromSet(messages, m)$$

Messages schemes and data structures

$$
\begin{array}{l}
createClinetReadRequest(key) \triangleq \\
\quad [mtype \mapsto ClientReadRequest, \\
\quad\ mkey \mapsto key, \\
\quad\ mhash \mapsto key]
\end{array}
$$

$$
\begin{array}{l}
createClinetWriteRequest(key, value) \triangleq \\
\quad [mtype \mapsto ClientWriteRequest, \\
\quad\ mkey \mapsto key, \\
\quad\ mvalue \mapsto value, \\
\quad\ mhash \mapsto key]
\end{array}
$$

When the switch forwards a read request, it adds:
$1 - logIndex$: The replica should $executereplicaVars$ this index
before serving the request
$2 - seqNum$: Which is used to ensure the linearizability
of the request response

$createInternalReadRequest(msg,\ KGroup,\ forwardTo) \triangleq$
$[mtype \mapsto InternalReadRequest,$
$mkey \mapsto msg.mkey,$
$mhash \mapsto msg.mhash,$
$msession \mapsto session,$
$mterm \mapsto switchTermId,$
$mleaderId \mapsto switchLeaderId,$
$mlogIndex \mapsto KGroup.logIndex,$
$mkGroupSeqNum \mapsto KGroup.seqNum,$
$msource \mapsto SwitchIp,$
$mdest \mapsto forwardTo]$

When the switch forwards a write request, it adds a
sequence number that is used to order write requests.
The leader processes the requests in order.

$createInternalWriteRequest(msg,\ KGroup) \triangleq$
$[mtype \mapsto InternalWriteRequest,$
$mkey \mapsto msg.mkey,$
$mvalue \mapsto msg.mvalue,$
$mhash \mapsto msg.mhash,$
$msession \mapsto session,$
$mterm \mapsto switchTermId,$
$mleaderId \mapsto switchLeaderId,$
$mkGroupSeqNum \mapsto KGroup.seqNum,$
$msource \mapsto SwitchIp,$
$mdest \mapsto switchLeaderId]$

$createReadResponse(m,\ i,\ leaderId,\ value,\ status,\ logIndex) \triangleq$
$[mtype \mapsto ReadResponse,$
$mkey \mapsto m.mkey,$
$mvalue \mapsto value,$
$mhash \mapsto m.mhash,$
$mstatus \mapsto status,$
$mlogIndex \mapsto logIndex,$
$mkGroupSeqNum \mapsto m.mkGroupSeqNum,$
$mterm \mapsto currentTerm[i],$
$mleaderId \mapsto leaderId,$
$mallLogs \mapsto log,$ for correctness check only
$mcommitIndex \mapsto commitIndex,$
$msession \mapsto m.msession,$

$msource \mapsto i,$
$mdest \mapsto SwitchIp]$

$createWriteResponse(i,\ logEntry,\ index,\ status,\ replicaIds) \triangleq$
$[mtype \mapsto WriteResponse,$
$mkey \mapsto logEntry.key,$
$mvalue \mapsto logEntry.value,$
$mhash \mapsto logEntry.hash,$
$mstatus \mapsto status,$
$mlogIndex \mapsto index,$
$mkGroupSeqNum \mapsto logEntry.seqNum,$
$msession \mapsto logEntry.switchId,$
$mreplicaIds \mapsto replicaIds,$
$mterm \mapsto currentTerm[i],$
$mallLogs \mapsto log,$ for correctness check only
$mcommitIndex \mapsto commitIndex,$
$msource \mapsto i,$
$mdest \mapsto SwitchIp]$

$createKGroup(leaderAcked,\ replicasIds,\ seqNum,\ logIndex) \triangleq$
$[leaderAcked \mapsto leaderAcked,$
$replicasIds \mapsto replicasIds,$
$seqNum \mapsto seqNum,$
$logIndex \mapsto logIndex]$

$createLogEntry(i,\ msg) \triangleq$
$[term \mapsto currentTerm[i],$
$key \mapsto msg.mkey,$
$value \mapsto msg.mvalue,$
$hash \mapsto msg.mhash,$
$seqNum \mapsto msg.mkGroupSeqNum,$
$switchId \mapsto msg.msession]$

$createResHistoryEntry(msg,\ tag) \triangleq$
$[msg \mapsto msg,$
$switchKGroupEntry \mapsto switchKGourpArray[getIndexFromHash(msg.mhash)],$
$tag \mapsto tag]$

Variables initialization

$InitSwitchVars \triangleq$
Initially the switch is inactive and

```
    KGroupArray is stable
    ∧ switchKGourpArray = [i ∈ 1 .. SwitchKGroupsNum
                            ↦ createKGroup(TRUE, {}, Nil, Nil)]
    ∧ switchState       = SwitchStateInactive
    ∧ switchTermId      = 0
    ∧ switchLeaderId    = Nil
    ∧ switchSeqNum      = 0
    ∧ session           = 0

InitMsgsSets ≜
    ∧ msgsClientSwitch      = {}
    ∧ msgsReplicasSwitch    = {}
    ∧ messages              = {}


InitReplicaVars ≜
    ∧ currentTerm = [i ∈ Replicas ↦ 1]
    ∧ state       = [i ∈ Replicas ↦ Follower]
    ∧ isActive    = [i ∈ Replicas ↦ ReplicaUpState]
    ∧ replicaSession = [i ∈ Replicas ↦ 0]
    ∧ log         = [i ∈ Replicas ↦ ⟨⟩]
    ∧ commitIndex = [i ∈ Replicas ↦ 0]

InitLeaderVars ≜
    ∧ nextIndex     = [i ∈ Replicas ↦ [j ∈ Replicas ↦ 1]]
    ∧ matchIndex    = [i ∈ Replicas ↦ [j ∈ Replicas ↦ 0]]
    ∧ replicaKGroups = [i ∈ Replicas ↦
                          [j ∈ 1 .. SwitchKGroupsNum ↦ 0]]

Init ≜ ∧ InitReplicaVars
       ∧ InitLeaderVars
       ∧ InitSwitchVars
       ∧ InitMsgsSets
        Needed to prove safety
       ∧ responsesToClient = {}
```

Variables actions

Client actions

client sends a read request to read key k

```
IssueReadRequest(key) ≜
    ∧ LET request ≜ createClinetReadRequest(key)
      IN  msgsClientSwitch' = AddToSet(msgsClientSwitch, request)
    ∧ UNCHANGED ⟨messages, replicaVars, leaderVars,
                 switchVars, msgsReplicasSwitch, responsesToClient⟩
```

```
 client issues a write request to update key k
IssueWriteRequest(key, value) ≜
    ∧ LET request ≜ createClinetWriteRequest(key, value)
      IN    msgsClientSwitch′ = AddToSet(msgsClientSwitch, request)
    ∧ UNCHANGED ⟨messages, replicaVars, leaderVars,
                 switchVars, msgsReplicasSwitch, responsesToClient⟩
```

Switch actions

```
 Switch state changes from Active to inactive
switchFails ≜
    ∧ switchState  =  SwitchStateActive
    ∧ switchState′ =  SwitchStateInactive
    ∧ UNCHANGED ⟨replicaVars, leaderVars, msgsVars, responsesToClient,
                 switchSeqNum, session, switchKGourpArray,
                 switchLeaderId, switchTermId⟩

 Switch handles a read request from a client.
 The request has a hash that is mapped to a
 KGroup. If the KGroup is stable, the request will
 be forwarded to one of the replcias, otherwise it
 will be forwarded to the leader.
SwitchHandleCleintRead(msg) ≜
  LET kGroup ≜ switchKGourpArray[getIndexFromHash(msg.mhash)]
      forwardTo  ≜ IF kGroup.leaderAcked ∧
                      kGroup.seqNum ≠ Nil
                   THEN ChooseRandomly({x ∈ kGroup.replicasIds :
                                        x ≠ switchLeaderId})
                   ELSE IF kGroup.leaderAcked ∧
                           kGroup.seqNum = Nil
                   THEN switchLeaderId
                   ELSE switchLeaderId
      internalMsg ≜ createInternalReadRequest(msg, kGroup, forwardTo)

  IN
    ∧ msgsReplicasSwitch′ = AddToSet(msgsReplicasSwitch, internalMsg)
    ∧ UNCHANGED ⟨replicaVars, leaderVars, switchVars,
                 responsesToClient, messages, msgsClientSwitch⟩

 Switch handles a write request from a client
SwitchHandleCleintWrite(msg) ≜
    Mark the KGroup associated with the key as unstable
  LET kGroup ≜ switchKGourpArray[getIndexFromHash(msg.mhash)]
```

```
        updatedKGroup  ≜ [kGroup EXCEPT !.leaderAcked = FALSE,
                                         !.seqNum = switchSeqNum + 1,
                                         !.replicasIds = {},
                                         !.logIndex = Nil]
        internalMsg    ≜ createInternalWriteRequest(msg, updatedKGroup)

    IN
        ∧ msgsReplicasSwitch' = AddToSet(msgsReplicasSwitch, internalMsg)
        ∧ switchKGourpArray' = [switchKGourpArray EXCEPT
                                ![getIndexFromHash(msg.mhash)] = updatedKGroup]
        ∧ switchSeqNum' =    switchSeqNum + 1  Increments the sequence number
        ∧ UNCHANGED ⟨replicaVars, leaderVars, responsesToClient,
                     messages, msgsClientSwitch, switchState,
                     switchTermId, switchLeaderId, session⟩


 Switch receives a read or write request from a client
SwitchReceiveFromClient ≜
    ∧ switchState = SwitchStateActive
    ∧ Cardinality(msgsClientSwitch) > 0
    ∧ LET msg  ≜ ChooseRandomly(msgsClientSwitch)
          type ≜ msg.mtype
      IN  ∨ ∧ type = ClientReadRequest
            ∧ SwitchHandleCleintRead(msg)
          ∨ ∧ type = ClientWriteRequest
            ∧ SwitchHandleCleintWrite(msg)


 Switch handles a read response
SwitchHandleReadResponse(msg) ≜
    ∧ ∨ ∧ msg.msource ≠ switchLeaderId  msg is from follower
        ∧ msg.mterm = switchTermId   msg.term = switch term
        ∧ msg.mstatus = TRUE    The operation was succeeded
        Map the key to a KGroup based on hash.
        The response will be sent to teh client if
        1 − msg.seqNum = KGroup.seqNum
        2 − KGroup is stable
        otherwise the request will be dropped
        ∧ LET KGroup ≜ switchKGourpArray[getIndexFromHash(msg.mhash)]
              isSeqOk    ≜ KGroup.seqNum = msg.mkGroupSeqNum

          IN  ∨ ∧ isSeqOk
                ∧ KGroup.leaderAcked
                ∧ responsesToClient' = AddToSet(
                                       responsesToClient,
                                       createResHistoryEntry(msg, Nil))
```

```
                    ∨ ∧ ¬isSeqOk
                      ∧ UNCHANGED ⟨responsesToClient⟩
                    ∨ ∧ ¬KGroup.leaderAcked
                      ∧ isSeqOk
                      ∧ UNCHANGED ⟨responsesToClient⟩
              ∨ ∧ msg.msource ≠ switchLeaderId
                ∧ msg.mstatus = FALSE  The key does not exist
                ∧ UNCHANGED ⟨responsesToClient⟩
             The operation succeeded and the response
             is from the current leader. Just forward it to the client
              ∨ ∧ msg.msource = switchLeaderId
                ∧ msg.mterm = switchTermId
                ∧ msg.mstatus = TRUE
                ∧ responsesToClient' = AddToSet(responsesToClient,
                                                createResHistoryEntry(msg, Nil))
              ∨ ∧ msg.msource = switchLeaderId
                ∧ msg.mstatus = FALSE  The key does not exist
                ∧ UNCHANGED ⟨responsesToClient⟩
           ∧ UNCHANGED ⟨replicaVars, leaderVars, switchVars, msgsVars⟩

 Switch handles a write response from a client.
 If the res.seqNum equals the KGroup.seqNum, the switch
 marks the KGroup as stable and forwards the response to
 the client. If the res.seqNum < KGroup.seqNum, the switch
 will not change the status of the KGroup, and will just
 forward the response to the clinet
SwitchHandleWriteResponse(msg) ≜
    ∧ ∨ ∧ msg.msource = switchLeaderId  Message from the leader
        ∧ msg.mterm   = switchTermId  Message.term = switch.term
        ∧ msg.mstatus = TRUE  The operation succeeded
        ∧ LET KGroup ≜ switchKGourpArray[getIndexFromHash(msg.mhash)]
              isSeqOk ≜ KGroup.seqNum = msg.mkGroupSeqNum
              updatedKGroup ≜ [KGroup EXCEPT
                                 !  .leaderAcked = TRUE,
                                 !  .replicasIds  = msg.mreplicaIds,
                                 !  .logIndex     = msg.mlogIndex]
             msg.seqNum = KGroup.seqNum
          IN ∨ ∧ isSeqOk
                Mark the KGroup as stable
                ∧ switchKGourpArray' = [switchKGourpArray EXCEPT
                                          ![getIndexFromHash(msg.mhash)] =
                                             updatedKGroup]
                ∧ responsesToClient' = AddToSet(responsesToClient,
                                                createResHistoryEntry(msg, Nil))
```

$msg.seqNum \,!= KGroup.seqNum$

$\lor \land \neg isSeqOk$
$\land responsesToClient' = AddToSet(responsesToClient,$
$createResHistoryEntry(msg,\ Nil))$
$\land$ UNCHANGED $\langle switchKGourpArray \rangle$

$\lor \land msg.msource \neq switchLeaderId$
$\land$ UNCHANGED $\langle switchKGourpArray,\ responsesToClient \rangle$

$\lor \land msg.mstatus = \text{FALSE}$
$\land$ UNCHANGED $\langle switchKGourpArray,\ responsesToClient \rangle$

$\land$ UNCHANGED $\langle replicaVars,\ leaderVars,\ msgsVars,$
$switchSeqNum,\ session,\ switchLeaderId,$
$switchTermId,\ switchState \rangle$

Switch receives a message from a replica

$SwitchReceiveFromReplica(msg) \triangleq$

msg term id is larger than switch term id
$\Rightarrow$ switch stops processing request by setting its status to inactive

$\lor \land switchState = SwitchStateActive$
$\land msg.msession = session$
$\land msg.mterm > switchTermId$
$\land switchState' = SwitchStateInactive$
$\land$ UNCHANGED $\langle replicaVars,\ leaderVars,\ msgsVars,$
$msgsClientSwitch,\ switchVars,\ responsesToClient \rangle$

msg is coming from an old leader
$\Rightarrow$ switch just ignore the message

$\lor \land switchState = SwitchStateActive$
$\land msg.msession = session$
$\land msg.mterm < switchTermId$
$\land$ UNCHANGED $\langle replicaVars,\ leaderVars,\ msgsVars,$
$responsesToClient,\ switchVars \rangle$

$msg.switchId$ does not match $switchId$
$\Rightarrow$ switch just ignore the message

$\lor \land switchState = SwitchStateActive$
$\land msg.msession \neq session$
$\land$ UNCHANGED $\langle replicaVars,\ leaderVars,\ msgsVars,$
$responsesToClient,\ switchVars \rangle$

Switch is active and the read response passes the safety check
$\Rightarrow$ switch processes the read response

$\lor \land switchState = SwitchStateActive$
$\land msg.msession = session$
$\land msg.mtype = ReadResponse$
$\land SwitchHandleReadResponse(msg)$

```
    Switch is active and the write response passes the safety check
    ⇒ switch processes the write response
  ∨ ∧ switchState = SwitchStateActive
    ∧ msg.msession = session
    ∧ msg.mtype = WriteResponse
    ∧ SwitchHandleWriteResponse(msg)
```

Replica actions

```
Replica i fails and stops processing msgs.
It loses everything but its currentTerm and log.
Stop(i) ≜
    ∧ isActive[i] = ReplicaUpState
    ∧ isActive'   = [isActive EXCEPT ![i] = ReplicaDownState]
    ∧ UNCHANGED ⟨leaderVars, msgsVars, switchVars, responsesToClient,
                 currentTerm, log,  commitIndex,
                 state, replicaSession⟩

Replica i becomes active. The replica stars as follower
Start(i) ≜
    ∧ isActive[i]     = ReplicaDownState
    ∧ isActive'       = [isActive EXCEPT ![i] = ReplicaUpState]
    ∧ state'          = [state EXCEPT ![i] = Follower]
    ∧ nextIndex'      = [nextIndex EXCEPT ![i] = [j ∈ Replicas ↦ 1]]
    ∧ matchIndex'     = [matchIndex EXCEPT ![i] = [j ∈ Replicas ↦ 0]]
    ∧ commitIndex'    = [commitIndex EXCEPT ![i] = 0]
    ∧ replicaKGroups' = [replicaKGroups EXCEPT ![i] =
                                  [j ∈ 1 .. SwitchKGroupsNum ↦ 0]]
    ∧ UNCHANGED ⟨msgsVars, switchVars, responsesToClient,
                 currentTerm, log, replicaSession⟩

Select a new leader if non of the replicas is a leader
This helper selects the new leader based on raft criteria.
That is, the node with the log with the highest term id and
longest log, if mutiple nodes have the same log id.
ElectLeader ≜
    ∧ LET runningReplicas ≜ {s ∈ Replicas : isActive[s] = ReplicaUpState}
          replicasWithNonEmptyLog ≜ {s ∈ runningReplicas : Len(log[s]) > 0}
          latestTerm ≜ Max({log[s][Len(log[s])].term :
                                  s ∈ replicasWithNonEmptyLog})
          replicasWithLatestTerm ≜ {s ∈ replicasWithNonEmptyLog :
                                    log[s][Len(log[s])].term = latestTerm}
          lengthOfLargestLog ≜ Max({Len(log[s]) :
                                    s ∈ replicasWithLatestTerm})
```

```
        replicasWithLongestLog ≜ {s ∈ replicasWithLatestTerm :
                                   Len(log[s]) = lengthOfLargestLog}
        newLeaderId ≜ IF   Cardinality(replicasWithLongestLog) ≥ 1
                      THEN CHOOSE i ∈ replicasWithLongestLog : TRUE
                      ELSE IF Cardinality(runningReplicas) > 0
                      THEN CHOOSE i ∈ runningReplicas : TRUE
                      ELSE Nil
        newTerm   ≜ IF newLeaderId ≠ Nil
                    THEN currentTerm[newLeaderId] + 1 ELSE Nil
        majority  ≜ IF newLeaderId ≠ Nil
                    THEN CHOOSE g ∈ Quorum : newLeaderId ∈ g
                    ELSE Nil
        newCurrentTerm ≜ [j ∈ Replicas ↦ IF j ∈ majority
                                          THEN newTerm
                                          ELSE currentTerm[j]]
    IN  ∧ Cardinality(runningReplicas) * 2 > Cardinality(Replicas)
        ∧ ∀ i ∈ runningReplicas : state[i] ∈ {Follower}
        ∧ state'      = [state EXCEPT ![newLeaderId] = Leader]
        ∧ nextIndex'  = [nextIndex EXCEPT ![newLeaderId] =
                        [j ∈ Replicas ↦ Len(log[newLeaderId]) + 1]]
        ∧ matchIndex' = [matchIndex EXCEPT ![newLeaderId] =
                        [j ∈ Replicas ↦ 0]]
        ∧ currentTerm' = newCurrentTerm
  ∧ UNCHANGED ⟨switchVars, msgsVars, responsesToClient,
               replicaKGroups, commitIndex, log,
               isActive, replicaSession⟩

 leader i populates the switch KGroup array
 with information about unstable KGroups
fillSwitchKGroup(i) ≜
  LET startIndex ≜ 1
      endIndex   ≜ Len(log[i])
       Scan the log and map each key in the log
       to its associated KGroup
      keysToKGroups ≜ [j ∈ startIndex .. endIndex ↦
                       getIndexFromHash(log[i][j].hash)]
       For each KGroup, find the last log entry
       that should be used to update the switch
       KGroup
      mapLogEntryToKGroupIndex ≜
        [j ∈ 1 .. SwitchKGroupsNum ↦
        IF Cardinality({k ∈ DOMAIN keysToKGroups : keysToKGroups[k] = j}) > 0
         THEN Max({k ∈ DOMAIN keysToKGroups : keysToKGroups[k] = j})
         ELSE Nil]
```

```
        A boolean array that indicates whether
        a log entry is committed or not
    leaderAckedFlag ≜ [j ∈ startIndex .. endIndex ↦
                          j ≤ commitIndex[i]]
        Get the set of replicas that
        acknowleged each log entry
    keysToReplicas  ≜ [j ∈ startIndex .. endIndex ↦
                          IF leaderAckedFlag[j]
                           THEN AgreeIndex(j, log, i)
                           ELSE {i}]
IN      Update the switch KGroup Array to match the leader's
        KGroup. If the leader do not have any writes for a
        KGroup, then the KGroup is stable
  ∧ switchKGourpArray' =
      [j ∈ 1 .. SwitchKGroupsNum ↦
      IF   mapLogEntryToKGroupIndex[j] ≠ Nil
       THEN createKGroup(leaderAckedFlag[mapLogEntryToKGroupIndex[j]],
                         keysToReplicas[mapLogEntryToKGroupIndex[j]],
                         0, mapLogEntryToKGroupIndex[j])
       ELSE createKGroup(TRUE, Replicas, Nil, Nil)]


 Leader i activates the switch
LeaderActivateSwitch(i) ≜
    ∧ state[i] = Leader  Replica is the leader
    ∧ isActive[i]  = ReplicaUpState  Replica is active
    ∧ switchState = SwitchStateInactive  Switch is inactive
    ∧ replicaSession' = [replicaSession EXCEPT ![i] = session + 1]
    ∧ session' = session + 1  Update the replica and switch sessions
    ∧ switchTermId' = currentTerm[i]  Update switch term number
    ∧ switchLeaderId' = i  Update leader id of the switch
    ∧ switchSeqNum' = 0  Reset the sequence number
    ∧ fillSwitchKGroup(i)  Populate switch KGroup array
    ∧ switchState' = SwitchStateActive  activate the switch
    ∧ UNCHANGED ⟨leaderVars, msgsVars, responsesToClient,
                 state, log, commitIndex, currentTerm,
                 isActive⟩


 Any RPC with a newer term causes the recipient to advance its term first.
UpdateTerm(i, j, m) ≜
    ∧ m.mterm > currentTerm[i]
    ∧ currentTerm'   = [currentTerm EXCEPT ![i] = m.mterm]
    ∧ state'         = [state       EXCEPT ![i]  = Follower]
      messages is unchanged so m can be processed further.
```

```
    ∧ UNCHANGED ⟨switchVars, msgsVars, leaderVars,
                 responsesToClient, isActive, log,
                 commitIndex, replicaSession⟩

 Responses with stale terms are ignored.
DropStaleResponse(i, j, m) ≜
    ∧ m.mterm < currentTerm[i]
    ∧ UNCHANGED ⟨replicaVars, leaderVars, switchVars,
                 responsesToClient, msgsVars⟩


 Replica i receives a read request (m) from a client.
ReplicaReceiveReadRequest(m, i) ≜

 The request received by the leader
∨ ∧ state[i] = Leader
   Check the message term numebr
   ∧ m.mterm = currentTerm[i]
   Get the index of the last committed entry
   ∧ LET committedEntries ≜ {j ∈ DOMAIN log[i] : ∧ log[i][j].key = m.mkey
                                                  ∧ j ≤ commitIndex[i]}
         lastEntryIndex ≜ IF Cardinality(committedEntries) > 0
                          THEN Max(committedEntries) ELSE Nil
         success ≜ IF lastEntryIndex = Nil THEN FALSE ELSE TRUE
         value ≜ IF success THEN log[i][lastEntryIndex].value ELSE Nil
     IN  ∧ msgsReplicasSwitch' = AddToSet(msgsReplicasSwitch,
                                  createReadResponse(m, i, getLeaderId,
                                  value, success, lastEntryIndex))
         ∧ UNCHANGED ⟨replicaVars, leaderVars, switchVars,
                      msgsClientSwitch, messages,
                      responsesToClient⟩

 The request received by a follower and msg.mlogIndex > 0
 i.e, the switch processed a write associated with the same
 KGroup of the key
∨ ∧ state[i] = Follower
   ∧ m.mterm = currentTerm[i]   msg.term = replica term
   ∧ m.mlogIndex ≤ Len(log[i])   The replica has the index
   ∧ m.mlogIndex > 0  Switch Kgroup entry is not empty
   Get the last committed log entry for the requested key
   ∧ LET logEntriesForKey ≜ entriesForKey(i, m.mkey)
         filteredEntries ≜ {j ∈ logEntriesForKey : j ≤ m.mlogIndex}
         lastEntryIndex ≜ IF Cardinality(filteredEntries) > 0
                          THEN Max(filteredEntries) ELSE Nil
         requestedEntry ≜ IF Cardinality(filteredEntries) > 0
                          THEN log[i][lastEntryIndex] ELSE Nil
```

```
        isCommitted ≜ m.mlogIndex ≤ commitIndex[i]
        success ≜ IF Cardinality(filteredEntries) > 0
                  THEN requestedEntry.key = m.mkey ELSE FALSE
        value ≜ IF success THEN requestedEntry.value ELSE Nil
    IN  ∧ msgsReplicasSwitch' = AddToSet(msgsReplicasSwitch,
                                  createReadResponse(m, i, getLeaderId,
                                      value, success, lastEntryIndex))
        ∧ ∨ ∧ isCommitted
            ∧ UNCHANGED ⟨commitIndex⟩
          ∨ ∧ ¬isCommitted
            ∧ success
            ∧ commitIndex' = [commitIndex EXCEPT ![i] = m.mlogIndex]
    ∧ UNCHANGED ⟨leaderVars, switchVars, responsesToClient,
                 log, state, currentTerm, isActive,
                 msgsClientSwitch, messages, replicaSession⟩

 The request received by a follower and msg.mlogIndex = −1.
 i.e., the switch did not process any write associated
∨ ∧ state[i] = Follower
  ∧ m.mterm = currentTerm[i]   msg.term = replica term
  ∧ m.mlogIndex = Nil          Switch Kgroup entry is empty
   Get the last committed log entry for the requested key
  ∧ LET committedEntries ≜ {j ∈ DOMAIN log[i] : ∧ log[i][j].key = m.mkey
                                                 ∧ j ≤ commitIndex[i]}
        lastEntryIndex ≜ IF Cardinality(committedEntries) > 0
                         THEN Max(committedEntries) ELSE Nil
        success ≜ IF lastEntryIndex = Nil THEN FALSE ELSE TRUE
        value ≜ IF success THEN log[i][lastEntryIndex].value ELSE Nil
    IN  ∧ msgsReplicasSwitch' = AddToSet(msgsReplicasSwitch,
                                  createReadResponse(m, i,
                                      getLeaderId, value, success,
                                      lastEntryIndex))

  ∧ UNCHANGED ⟨replicaVars, leaderVars, switchVars,
               responsesToClient, msgsClientSwitch,
               messages⟩


 Leader i receives a write request.
ReplicaReceiveWriteRequest(m, i) ≜
    Safety checks
  ∨ ∧ m.mterm = currentTerm[i]   msg.term = replica.term
    ∧ m.mleaderId = i            The leaderId field in the message is the replica
    ∧ state[i] = Leader          The replica is the leader
    ∧ m.msession = replicaSession[i]   msg.session = replica.session
     Get the highest sequence number received fot his KGroup
```

```
        ∧ LET latestSeenSeqNum ≜ replicaKGroups[i][getIndexFromHash(m.mhash)]
              entry ≜ createLogEntry(i, m)
              newLog ≜ Append(log[i], entry)
               msg.seqNum > KGroup.seqNum
          IN  ∨ ∧ m.mkGroupSeqNum > latestSeenSeqNum
                 Update the KGroup seqNum
                ∧ replicaKGroups' = [replicaKGroups EXCEPT ![i] =
                                       [@ EXCEPT ![getIndexFromHash(m.mhash)] =
                                       m.mkGroupSeqNum]]
                ∧ log' = [log EXCEPT ![i] = newLog]  Append to log
                ∧ UNCHANGED ⟨switchVars, msgsVars, responsesToClient,
                             nextIndex, matchIndex, commitIndex,
                             replicaSession, state, currentTerm, isActive⟩
              ∨ ∧ m.mkGroupSeqNum ≤ latestSeenSeqNum
                ∧ UNCHANGED ⟨replicaVars, leaderVars, switchVars,
                             responsesToClient, msgsVars⟩
    ∨ ∧ ∨ m.mterm      ≠ currentTerm[i]
        ∨ m.mleaderId  ≠ i
        ∨ state[i]     ≠ Leader
        ∨ m.msession ≠ replicaSession[i]
      ∧ UNCHANGED ⟨replicaVars, leaderVars, switchVars,
                   responsesToClient, msgsVars⟩
```

Leader i sends follower j an *AppendEntries* request containing up to 1 entry. While implementations may want to send more than 1 at a time, this spec uses just 1 because it minimizes atomic regions without loss of generality.

```
AppendEntries(i, j) ≜
    ∧ i ≠ j  Avoid sending to itself
    ∧ state[i] = Leader  The sender is the leader
    ∧ isActive[i] = ReplicaUpState  The sender is active
     Get the index of the next entry that must
     be sent to the follower j
    ∧ LET prevLogIndex ≜ nextIndex[i][j] − 1
          prevLogTerm ≜ IF prevLogIndex > 0 THEN
                           log[i][prevLogIndex].term
                        ELSE
                           0
       Send up to 1 entry, constrained by the end of the log.
      lastEntry ≜ Min({Len(log[i]), nextIndex[i][j]})
      entries ≜ SubSeq(log[i], nextIndex[i][j], lastEntry)
      m ≜       [mtype          ↦ AppendEntriesRequest,
                 mterm          ↦ currentTerm[i],
                 mprevLogIndex  ↦ prevLogIndex,
                 mprevLogTerm   ↦ prevLogTerm,
```

```
                    mentries       ↦ entries,
                    mlog is used as a history variable for the proof.
                    It would not exist in a real implementation.
                    mlog           ↦ log[i],
                    mcommitIndex   ↦ Min({commitIndex[i], lastEntry}),
                    msource        ↦ i,
                    mdest          ↦ j]
  IN    ∧ Len(entries) > 0
        ∧ Send(m)
  ∧ UNCHANGED ⟨replicaVars, leaderVars, switchVars,
               msgsClientSwitch, msgsReplicasSwitch,
               responsesToClient⟩


Leader i advances its commitIndex.
This is done as a separate step from handling AppendEntries responses,
in part to minimize atomic regions, and in part so that leaders of
single-server clusters are able to mark entries committed.

AdvanceCommitIndex(i) ≜
  ∧ state[i] = Leader
  ∧ isActive[i] = ReplicaUpState
  ∧ LET The set of replicas that agree up through an index.
        Agree(index) ≜ {i} ∪ {k ∈ Replicas :
                              matchIndex[i][k] ≥ index}
        The maximum indexes for which a quorum agrees
        agreeIndexes ≜ {index ∈ 1 .. Len(log[i]) :
                        Agree(index) ∈ Quorum}
        Entries that are replicated to majorities
        but not yet committed
        IndicesToCommit ≜ {index ∈ agreeIndexes : index > commitIndex[i]}
        majoritiesPerIndex ≜ [index ∈ agreeIndexes ↦ Agree(index)]
        New value for commitIndex'[i]
        newCommitIndex ≜
          IF ∧ agreeIndexes ≠ {}
             ∧ log[i][Max(agreeIndexes)].term = currentTerm[i]
           THEN Max(agreeIndexes)
           ELSE commitIndex[i]
    IN  ∧ newCommitIndex > commitIndex[i]
        For each entry that will be committed, send a write
        response to clients
        ∧ LET indices ≜ commitIndex[i] + 1 .. newCommitIndex
              msgs ≜ [x ∈ indices ↦
                      createWriteResponse(i, log[i][x],
```

```
                                                    x, TRUE, majoritiesPerIndex[x])]
            msgsAsSet ≜ {createWriteResponse(i, log[i][x],
                                                    x, TRUE, majoritiesPerIndex[x])
                                                      : x ∈ indices}

    IN  ∧ msgsReplicasSwitch' = msgsReplicasSwitch ∪ msgsAsSet
        Update the commit index at the replica
        ∧ commitIndex' = [commitIndex EXCEPT ![i] = newCommitIndex]
        ∧ UNCHANGED ⟨leaderVars, switchVars, responsesToClient,
                     messages, msgsClientSwitch, log, currentTerm,
                     isActive, replicaSession, state⟩

Replica i receives an AppendEntries request from Replica j.
This just handles m.entries of length 0 or 1, but
implementations could safely accept more by treating
them the same as multiple independent requests of 1 entry.
HandleAppendEntriesRequest(i, j, m) ≜
  LET logOk ≜ ∨ m.mprevLogIndex = 0
              ∨ ∧ m.mprevLogIndex > 0
                ∧ m.mprevLogIndex ≤ Len(log[i])
                ∧ m.mprevLogTerm = log[i][m.mprevLogIndex].term
  IN  ∧ m.mterm ≤ currentTerm[i]
      ∧ ∨ ∧ reject request
          ∨ m.mterm < currentTerm[i]
          ∨ ∧ m.mterm = currentTerm[i]
            ∧ state[i] = Follower
            ∧ ¬logOk
        ∧ Send([mtype      ↦ AppendEntriesResponse,
                mterm      ↦ currentTerm[i],
                msuccess   ↦ FALSE,
                mmatchIndex ↦ 0,
                msource    ↦ i,
                mdest      ↦ j])
        ∧ UNCHANGED ⟨replicaVars⟩
        process the request
      ∨ ∧ m.mterm = currentTerm[i]
        ∧ state[i] = Follower
        ∧ logOk
        ∧ LET index ≜ m.mprevLogIndex + 1
          IN  ∨ already done with request
                ∧ ∨ m.mentries = ⟨⟩
                  ∨ ∧ Len(log[i]) ≥ index
                    ∧ m.mentries ≠ ⟨⟩
                    ∧ log[i][index].term = m.mentries[1].term
                This could make our commitIndex decrease (for
```

```
                example if we process an old, duplicated request),
                but that doesn't really affect anything.
            ∧ commitIndex' = [commitIndex EXCEPT ![i] =
                                 m.mcommitIndex]
            ∧ Send([mtype       ↦ AppendEntriesResponse,
                    mterm       ↦ currentTerm[i],
                    msuccess    ↦ TRUE,
                    mmatchIndex ↦ m.mprevLogIndex +
                                  Len(m.mentries),
                    msource     ↦ i,
                    mdest       ↦ j])
            ∧ UNCHANGED ⟨log⟩
        ∨ conflict: remove 1 entry
            ∧ m.mentries ≠ ⟨⟩
            ∧ Len(log[i]) ≥ index
            ∧ log[i][index].term ≠ m.mentries[1].term
            ∧ LET new ≜ [index2 ∈ 1 .. (Len(log[i]) − 1) ↦
                           log[i][index2]]
              IN  log' = [log EXCEPT ![i] = new]

            ∧ Send([mtype       ↦ AppendEntriesResponse,
                    mterm       ↦ currentTerm[i],
                    msuccess    ↦ FALSE,
                    mmatchIndex ↦ 0,
                    msource     ↦ i,
                    mdest       ↦ j])
            ∧ UNCHANGED ⟨commitIndex⟩
        ∨ no conflict: append entry
            ∧ m.mentries ≠ ⟨⟩
            ∧ Len(log[i]) = m.mprevLogIndex
            ∧ log' = [log EXCEPT ![i] =
                        Append(log[i], m.mentries[1])]
            ∧ Send([mtype       ↦ AppendEntriesResponse,
                    mterm       ↦ currentTerm[i],
                    msuccess    ↦ TRUE,
                    mmatchIndex ↦ m.mprevLogIndex +
                                  Len(m.mentries),
                    msource     ↦ i,
                    mdest       ↦ j])
            ∧ UNCHANGED ⟨commitIndex⟩
    ∧ UNCHANGED ⟨leaderVars, switchVars, responsesToClient,
                 msgsClientSwitch, msgsReplicasSwitch,
                 isActive, currentTerm, state, replicaSession⟩
```

Replica i receives an *AppendEntries* response from Replica j

```
HandleAppendEntriesResponse(i, j, m) ≜
    ∧ m.mterm = currentTerm[i]
    ∧ ∨ ∧ m.msuccess  successful
        ∧ nextIndex'  = [nextIndex  EXCEPT ![i][j]  = m.mmatchIndex + 1]
        ∧ matchIndex' = [matchIndex EXCEPT ![i][j] = m.mmatchIndex]
      ∨ ∧ ¬m.msuccess  not successful
        ∧ nextIndex' = [nextIndex EXCEPT ![i][j] =
                            Max({nextIndex[i][j] − 1, 1})]
        ∧ UNCHANGED ⟨matchIndex⟩
    ∧ Discard(m)
    ∧ UNCHANGED ⟨replicaVars, switchVars, responsesToClient,
                 msgsClientSwitch, msgsReplicasSwitch, replicaKGroups⟩


 process a message. The message will be processed
 by its recipient based on its type
Receive(m) ≜
    LET i ≜ m.mdest
        j ≜ m.msource
    IN   Any RPC with a newer term causes the recipient to advance
         its term first. Responses with stale terms are ignored.
    ∧   isActive[i] = ReplicaUpState
    ∧   ∨ ∧ m.mtype ∈ {AppendEntriesRequest, AppendEntriesResponse}
          ∧ UpdateTerm(i, j, m)
         Append entry request from replica j to replica i
        ∨ ∧ m.mtype = AppendEntriesRequest
          ∧ HandleAppendEntriesRequest(i, j, m)
         Append entry response from replica j to replica i
        ∨ ∧ m.mtype = AppendEntriesResponse
          ∧ ∨ DropStaleResponse(i, j, m)
            ∨ HandleAppendEntriesResponse(i, j, m)

         Read request from the switch to replica i
        ∨ ∧ m.mtype = InternalReadRequest
          ∧ ReplicaReceiveReadRequest(m, i)
         Write request from the switch to replica i
        ∨ ∧ m.mtype = InternalWriteRequest
          ∧ ReplicaReceiveWriteRequest(m, i)

ReplicasReceiveRaftInternalMsgs ≜
    ∧ ∃ m ∈ messages : Receive(m)

ReplicasReceiveFromSwitch ≜
    ∧ ∃ m ∈ msgsReplicasSwitch :
            ∧ m.mdest ≠ SwitchIp ∧ Receive(m)
```

Defines how all *variables*(*Replicas*, *Client*, *Switch*)
may transition.

$$
\begin{aligned}
&Next \triangleq \\
&\quad \text{client transitions} \\
&\quad \land \lor \exists\, key \in KeySpace : IssueReadRequest(key) \\
&\quad\quad \lor \exists\, key \in KeySpace : \exists\, value \in ValueSpace : \\
&\quad\quad\quad\quad IssueWriteRequest(key,\, value) \\
&\quad \text{Switch transitions} \\
&\quad \lor switchFails \\
&\quad \lor SwitchReceiveFromClient \\
&\quad \lor \land Cardinality(msgsReplicasSwitch) > 0 \\
&\quad\quad \land \text{LET } messagesToSwitch \triangleq \{x \in msgsReplicasSwitch : x.mdest = SwitchIp\} \\
&\quad\quad\quad \text{IN} \quad \land \exists\, msg \in messagesToSwitch : SwitchReceiveFromReplica(msg) \\
&\quad \text{Replica transitions} \\
&\quad \lor \exists\, i \in Replicas : Stop(i) \\
&\quad \lor \exists\, i \in Replicas : Start(i) \\
&\quad \lor ElectLeader \\
&\quad \lor ReplicasReceiveRaftInternalMsgs \\
&\quad \lor ReplicasReceiveFromSwitch \\
&\quad \text{Leader transitions} \\
&\quad \lor \exists\, i \in Replicas : LeaderActivateSwitch(i) \\
&\quad \lor \exists\, i, j \in Replicas : AppendEntries(i, j) \\
&\quad \lor \exists\, i \in Replicas \quad : AdvanceCommitIndex(i)
\end{aligned}
$$

Safety Invariants

In the following statements, each response has the
logs of all replicas at the time it was served
by a replica (This is only for safety check and
should not be the case for real implementations)

Invariant that defines that read responses that
passes the switch to the client are linearizable.
Check the logs of all replica to get the last committed
log index that update the requested key. The returned value
"*msg.mvalue*" should equal the value in that *log* index.

$$
\begin{aligned}
&isForwardedToClientReadSafe(response) \triangleq \\
&\quad \text{LET } msg \triangleq response.msg \\
&\quad\quad logEntriesForKey \triangleq entriesForKey(msg.mleaderId,\, msg.mkey)
\end{aligned}
$$

```
        committedEntries ≜ {j ∈ logEntriesForKey :
                            AgreeIndex(j, msg.mallLogs, msg.mleaderId)
                             ∈ Quorum ∧ j ≤ msg.mlogIndex}
        lastCommittedIndex ≜ indexOfLastEntry(committedEntries)
        lastEntry ≜ log[msg.mleaderId][lastCommittedIndex]
    IN
        ∨ ∧ msg.mstatus = TRUE
          ∧ lastCommittedIndex = msg.mlogIndex
          ∧ lastEntry.key      = msg.mkey
          ∧ lastEntry.value    = msg.mvalue

 Invariant that defines that write responses that
 passes the switch to the client are committed
 on majority of the replicas.
isForwardedToClientWriteCorrect(response) ≜
    LET msg ≜ response.msg
        isOnMajority ≜ AgreeIndex(msg.mlogIndex,
                                  msg.mallLogs, msg.msource) ∈ Quorum
    IN  isOnMajority

 Invariant that defines the correctness
 of all responses forwarded to the client
InvResponsesToClientCorrectness ≜
    ∀ res ∈ responsesToClient :
      ∨ ∧ res.msg.mtype = ReadResponse
        ∧ isForwardedToClientReadSafe(res)
      ∨ ∧ res.msg.mtype = WriteResponse
        ∧ isForwardedToClientWriteCorrect(res)

 Invariant that defines that no two entries have
 the same sequence number or log index.
InvSwitchRegisterCorrectness ≜
    ∧ ∀ i, j ∈ 1 .. SwitchKGroupsNum :
        ∨ ∧ i ≠ j
          ∧ switchKGourpArray[i].leaderAcked = TRUE
          ∧ switchKGourpArray[j].leaderAcked = TRUE
          ∧ switchKGourpArray[i].seqNum ≠ switchKGourpArray[j].seqNum
          ∧ switchKGourpArray[i].logIndex ≠ switchKGourpArray[j].logIndex
        ∨ i = j
        ∨ switchKGourpArray[i].leaderAcked = FALSE
        ∨ switchKGourpArray[j].leaderAcked = FALSE


 invariant that defines that no two leaders exist
 at the same time
InvLeaderElectionSafety ≜
```

$$
\begin{array}{ll}
\text{LET} & runningReplicas \triangleq \{i \in Replicas : isActive[i] = ReplicaUpState\} \\
& leaders \triangleq \{i \in runningReplicas : state[i] = Leader\} \\
\text{IN} & Cardinality(leaders) \leq 1
\end{array}
$$