

R2 protocol, host and storage refinement

This is a compositional argument, with executable models and implementation regressions for its boundaries. The completion evidence below closes R2 under these explicit assumptions. Final-candidate integration remains R7.

Assumptions and supported failure model

The fixed three-voter membership uses intersecting majorities. Voters are non-Byzantine, authenticate peers, preserve acceptor identity and durable promises, and run the same enforced SQL/extension policy. Storage honors successful FULL commits and required file/directory synchronization. Unknown storage failures stop a voter; a lost acceptor is never recreated under its previous identity. R4's fresh-namespace replacement procedure applies after irrecoverable disk loss.

Safety does not require bounded message delay. Liveness requires a surviving majority, eventual successful retransmission, fair service of admitted work, bounded application/storage work, and an eventually non-preempted phase-one/phase-two exchange. On the qualified LAN, communication and durable processing must fit inside the configured election interval. The current default is ten 100 ms ticks. This is not a claim that fixed timeouts guarantee progress at every possible eventual network delay, or that clients are wait-free during perpetual overload. All such limits remain explicit in R5/R6 measurements.

Per-decree agreement

The complete pinned paxos-odin library owns ballot selection, acceptors, learners, recovery and rotating ownership. SQLodin does not duplicate its consensus logic. Slot ownership is $(slot - 1) \bmod member_count$; only that owner may use the slot's round-zero ballot. Revocation uses a higher ballot with a distinct proposer ID. An acceptor votes only at or above its durable promise, and a proposer selects the highest accepted value reported by its phase-one quorum. A ballot does not issue two different values for the same decree.

For any previously chosen value, the later phase-one quorum intersects its voting quorum. By induction on higher ballots, highest-vote selection carries that value forward. The round-zero owner is the unique initial proposer for its decree, so it is the base case of the same argument, not a bypass around phase one at later ballots. Per-slot bounded promises revoke only the intended range. The OwnedSlot model checks conflicting owners/revokers and the missing-durability counterexample; the finite ballot bound is not itself an unbounded proof. The quorum-intersection induction supplies the general argument under ballot uniqueness and monotonicity.

Code boundaries are `ownership_ballot`, `propose_owned`, `start_revocation`, `promise_bounded`, `on_accept`, `maybe_resolve_chunk` and `resolve_chunk` in the pinned dependency. The SQLodin adapter enables rotating ownership and forwards the complete values and effects. Wire/codec tests check ownership of copied payloads.

Durable effects and application prefixes

`durable.persist` and `commit_group_journal` write promises, votes and decisions before `effects_confirm_writes_durable` exposes their dependent messages. A local acceptor's loopback messages cross the same durable transition boundary. Journal grouping changes the number of FULL commits, not the ordering of the barrier; a service turn stages all of its transitions behind one barrier (SOD 0005 M1). The host may step an owner's round-zero Accept as a Commit when it carries the host no-op, or, with a write quorum of two, when this voter's own vote for the same value is in the same group (SOD 0005 M2/M5; `OwnedSkip.tla`, `OwnedSkipProof.tla`). M2 no-ops can be learned before an actual voting quorum exists. Thus the old DurableHistoryProof premise that every chosen record has a voting quorum is not literally satisfied by these synthetic Commit records.

OwnedSkip's no-op determinacy argument supplies the extra learner justification; state-changing values still need a durable voting quorum. The composed argument remains conditional, not a new machine-checked refinement of the entire host. Replay folds the durable records before the node resumes accepting requests.

Chosen slots apply contiguously. Application data, request fences, outcomes, revision and watermark commit together; acknowledgements require matching chosen and applied request evidence. `specs/grouped-sql.type` establishes the application group/reference relation. `specs/transaction-order.type` establishes fresh reads and optimistic transaction order. `specs/session-retirement.type` establishes retry fencing across bounded session reclamation. SQL transition determinism remains R1.1's premise, not something inferred from a consensus proof.

The checked DurableHistoryProof supplies the unbounded stable/released/chosen/ applied/ acknowledged induction. PrefixRecoveryProof adds arbitrary natural-numbered slots, contiguous committed prefixes, an image cut, retained suffix, trimming and recovery readiness. Its invariant states that every committed slot is chosen, the entire suffix above the certified cut is retained, and readiness implies replay reached the committed prefix. Acknowledged prefixes remain covered even when a crash resets the replay cursor to the image cut. The 36 discharged proof obligations use no omitted steps. The atomic published-image premise is discharged at the implementation seam by certificate validation, generation publication and their crash/model checks; this lemma does not prove image bytes by itself.

Rotating-window progress

For a finite offered frontier, consider the lowest unapplied slot. Its live owner either proposes pending work or emits an idle skip after observing the frontier. A missing owner's slot is recovered by a bounded phase-one exchange. Under the stated eventual exchange assumption, a majority's durable votes choose it. Learning and application advance the contiguous prefix; only that applied floor allows ring-cell reuse. Induction on the remaining distance reaches the frontier. Reapplying this argument to later frontiers establishes continuing prefix progress; bounded windows limit in-flight work rather than the lifetime log length.

RotatingWindow makes these dependencies executable. It permits independent delivery orders, deferred learning, different applied/freed frontiers and multiple owners. Durable voting and safe recovery offers are abstracted from the decree argument above. Weak fairness represents eventual successful retransmission, not zero packet loss. It does not separately model every revocation ballot or prove starvation freedom for an individual request under endless interference. The host retains a displaced request's identity; the pinned library prioritizes bounded resubmissions, and native tests verify retry/catch-up behavior.

Removing idle skips, absent-owner recovery or floor release must violate progress. Removing ring admission must overwrite a still-held slot. Each is a required negative control, matched to `owned_progress_fills_gaps_without_time`, `ownership_revokes_a_crashed_owner`, `window/floor continuation tests`, and `window_follower_refuses_slots_beyond_its_window` in the dependency. The host's `test_idle_owners_finish_without_timer_ticks` asserts completion with election and stall counters still zero, directly checking the old healthy-path timer penalty. Timer-driven failure detection remains separate from `durable.progress`.

Recovery and retirement mapping

Authenticated `catch_up` repeatedly requests bounded chunks without waiting for new client traffic. A lagging voter cannot block a healthy quorum's packet queues. Beyond retained history, snapshot transfer verifies the complete image/key and chosen certificate. Installation copies only application

state from the donor; the recipient preserves its own promises, votes, reserved IDs and accepted/chosen suffix. Catalog publication transfers the active store only after durable private construction, preserves displaced read/request fences, and poisons ambiguous old state. GenerationCatalog and the publication crash matrix cover this mapping.

Generation/image retirement preserves the active image and exact predecessor, and never discards an unsealed receipt without a chosen successor. Their ownership inventories and sync-before-forget rules map to GenerationRetirement and ImageRetirement. RestoredGenesis covers fresh-namespace bootstrap after fencing the old group; it is not same-identity acceptor replacement. The retained-history quota leaves transition reserves and fails closed on invalid inventory. These mechanisms discharge storage premises; they do not change the quorum algorithm.

Finite TLC runs, unbounded seam induction and native fault tests supply different kinds of evidence. None alone establishes end-to-end production readiness. Final-source integration and three-machine qualification remain R7.

Completion evidence

`formal-rotating-linux-v3.json` checks 737,901 concurrent-producer states plus the repeated-window/absent-voter cases and four required negative controls. In the all-producers/no-failure configuration, highest-seen does not affect any enabled transition or property; normalizing it removes equivalent observation states. The skew/failure configurations retain independent observations. Earlier unreduced runs timed out and remain recorded as incomplete, with their original model.

`formal-prefix-proof-linux-v3.json` discharges all 36 prefix induction obligations; `formal-proofs.json` retains the 12 durable-history obligations. The initial prefix proof attempts failed and remain archived. No omitted obligations are accepted. `consensus-policy8-linux-v3.json` records 81 passing pinned-core tests and 180,000 seeded SQLodin fault steps. `majority-policy8-linux.json` records native writes/reads through both survivors with each voter absent and recovery/restart. The policy-8 progress reports record zero-tick healthy completion and multi-window catch-up. Current live-generation evidence supplies certified-snapshot recovery. These reports are under `benchmarks/results/verification-20260924/`; R2 is complete and the remaining contract criteria are unchanged.