

GUJARAT POLICE INNOVATION CHALLENGE 2026

Sentinel

Evidence-First Vehicle

Journey System

High-Level Design

MANDATORY MODEL 1 + MODEL 2

- Dynamic registry and verified GIS
- Consume-only heterogeneous feed integration
- Exact multi-frame ANPR with abstention
- Watchlist alerts and PTS-aware route evidence
- Regional cells for statewide scale

Document status: submission candidate, evidence snapshot 31 August 2026

Selected architecture: mandatory Model 1 plus Model 2, with evidence-first route reconstruction

1. Executive decision

Sentinel answers one operational question: where was the designated vehicle observed, in what defensible order, and what source evidence supports each point?

The system combines the compulsory centralized camera registry and GIS foundation with direct, consume-only integration to existing departmental feeds. It centralizes camera metadata, sightings, alerts, route evidence, policy, and audit records. It does not require departments to replace their VMS or centralize all video.

This design deliberately rejects a statewide central-video copy as the default. At the published planning target of about 80,000 cameras, centralizing 1 to 2 Mbit/s streams would move 80 to 160 Gbit/s and create approximately 0.864 to 1.728 PB of raw video per day before replication. Regional analytics plus centralized searchable metadata gives the required investigation outcome with less bandwidth, storage, latency, and failure surface.

2. Success contract

An authorized officer enters a vehicle registration, receives the correct location-wise route quickly, verifies every material sighting against source evidence, and can act without being misled by false certainty.

The system succeeds only when it demonstrates:

- Dynamic onboarding from the official camera catalogue without hard-coded IDs or counts.
- A unified, demand-driven viewer across at least two provenance-confirmed source systems.
- Exact, multi-frame plate evidence with abstention on weak or conflicting reads.
- Automatic representative-watchlist alerts with duplicate suppression.
- A timestamped route whose observed points, uncertain order, and coverage gaps are visibly distinct.
- Consume-only operation that leaves departmental VMS and storage unaffected.
- Measured performance and a credible regional-cell path toward approximately 80,000 cameras.

3. Requirement and evidence boundary

Claim	Current evidence	Submission gate
Model 1 registry	Working local API and portal; 30-camera snapshot; API, manual and bulk onboarding; GIS gap behavior; health, maintenance, export and audit	Record canonical browser flow and retain build manifest
Unified viewer	WHEP-first, native-HLS fallback; operator-selected sessions capped at two; browser lifecycle verified	Successful simultaneous viewing from two confirmed source systems
Live government analytics	Complete 30-camera launcher decoded frames from 14 official cameras and completed 32 model jobs with zero model failures; transport remained degraded	Accepted readable live track with retained report
Watchlist and route	Exact-only versioned matcher, alerts, decisions, route ordering and reports work end to end	Run designated-vehicle rehearsal from live/recorded evidence
Statewide scale	80,000-record metadata benchmark and transparent video sensitivity model	Measure one mixed-codec regional inference cell before extrapolating compute

No model confidence is reported as accuracy. No camera location, department, source system, or route segment is inferred silently.

4. Logical architecture



```
Department cameras and existing VMS/NVR
-> catalogue adapter and feed supervisor
-> one paced decode session per active camera
-> bounded in-memory frame fan-out
-> crash-isolated persistent inference worker
-> one loaded RF-DETR/OIM/CCT scout bundle
-> scout detector and vehicle tracker
-> plate localization and crop-quality gate
-> fast OCR scout
-> one host-shared, serialized persistent Awiros verifier
-> exact multi-frame track consensus
-> sighting and evidence record
-> versioned watchlist correlation
-> deduplicated proposed alert
-> PTS-aware journey reconstruction
-> operator viewer, route timeline, evidence and audit

State control plane
<- regional camera health and coverage
<- searchable sighting metadata and selected policy-approved evidence
-> registry policy, watchlists, case scope and model/build manifest
```

Pixel processing remains regional or departmental where policy permits. The state layer holds the minimum searchable metadata and selected evidence needed for the investigation.

5. Component responsibilities

Component	Responsibility	Failure behavior
Catalogue adapter	Treat GET /api/ingest as the dynamic contract; preserve unknown fields; sanitize secrets	Retain last known safe registry; mark freshness and missing cameras
Registry and GIS	Camera identity, ownership metadata, source-system proof, declared/measured properties, health, maintenance and verified coordinates	Unknown remains explicit; no coordinate or ownership guessing
Feed supervisor	RTSP/TCP inference session, WHEP/HLS preview references, reconnect, pacing and codec/session health	Bounded exponential backoff; visible unavailable/degraded state
Decode and fan-out	Decode once, attach source PTS/timebase/session/loop state, feed bounded consumers	Drop stale work before memory or latency grows without bound
Persistent inference fleet	Each measured cell owns one exact RF-DETR/OIM/CCT scout bundle; all cells on one host share one serialized, lazy Awiros verifier	Specialist failure suppresses authority across that host; never reload per frame or silently change artifacts
Adaptive analytics	Scout, engage, pursuit and confirm policies; detection, tracking, plate localization and OCR	Preserve minimum service; abstain when evidence is unsupported; verifier failure suppresses sightings
Sighting service	Store exact candidates, evidence digest, model/build identity and temporal provenance	Reject duplicate/conflicting identities and unknown cameras
Watchlist engine	Versioned exact correlation, validity window, deduplication and proposed alerts	Fuzzy/conflicting/expired/paused records never alert automatically
Journey engine	Vehicle search, PTS-aware ordering, uncertainty, coverage gaps and evidence report	Partial or uncertain route is labeled; no invented interpolation
Operator application	Registry, viewer, alerts, route timeline, evidence review and decisions	Fail closed on role errors; show unavailable and uncertainty states
Audit and policy	Append-only metadata, search, watchlist, export and decision history	Corrections append; prior state remains traceable

6. Heterogeneous integration strategy

The government adapter isolates organizer and vendor-specific payloads. The canonical internal contracts remain unchanged when a source connector is replaced.

6.1 Discovery

- Fetch every current camera from the catalogue; never derive URLs from ID patterns.
- Retain raw unknown fields only after credential and URL-query sanitization.
- Separate source-declared codec, resolution, frame rate and bitrate from measured observations.
- Store department and source system independently. A source-system claim requires a provenance reference and explicit confirmed status.

6.2 Media

- RTSP over TCP is the authoritative inference transport when reachable.
- WHEP is the first browser-preview transport because it targets low latency.

- HLS is a dashboard and restricted-network fallback, not the inference clock.

If RTSP/TCP is unreachable, the bounded sequential HLS inference adapter is available only as an explicitly degraded fallback: live-edge latency is visible, complete segments are deduplicated and discarded after decode, and the warm model/association consumer is reused across segments. It does not replace the official RTSP/TCP path or alter OCR/authority thresholds.

- ONVIF, vendor SDK and departmental VMS APIs are connector options behind the same adapter boundary; none are claimed live without supplied access.
- Existing VMS/NVR storage remains authoritative for raw retention unless policy explicitly changes.

6.3 Timing and discontinuity

- Use source PTS, RTP or buffer PTS; never declared FPS or arrival time.
- Store source timebase, scenario time, ingest time, session, loop, uncertainty and timing quality separately.
- Treat initial buffered GOP acceleration as startup, not vehicle speed.
- A restart, loop or backward PTS creates a new boundary. Cross-boundary ordering abstains unless a trustworthy program-time anchor exists.
- Queue freshness alone uses a process-local monotonic clock. Independent

camera PTS domains are never compared for admission or capacity decisions.

6.4 Bounded feed-control runtime

`multi_camera_supervisor.py` owns admitted frame payloads under the scheduler's hard queue/in-flight caps. Coalescing, eviction, health loss, reconnect, loop, completion, and stale-session rejection all release payload ownership. A separate per-camera `FeedSession` applies the documented 2/4/8/16/30-second backoff, creates explicit reconnect sessions, increments known loop indices, and marks missing PTS, timebase changes, and unexplained rewinds without changing camera timestamps. The first runtime uses one model slot until parallel model execution is measured safe. Decoder callbacks retain exactly one latest frame per camera with coordinator-monotonic arrival time; the owner expires stale held frames before `FeedSession.offer` and admits only on an empty scheduler queue. Admission selects one deterministic round-robin SCOUT per due slot. At most one routine urgent frame shares every fourth scout; designated-target work has a separate three-scout bound. No urgent work consumes the between-scout reserve. Active target-interest cameras rank first when their bound is due, so an investigation cannot be hidden by a busier camera. Loop markers are sticky across mailbox replacement and cleared at failure/reconnect/token boundaries. Scout admission uses a system-owned interval derived from the ten-second deadline and hard camera ceiling rather than declared FPS, a simulation duration, or an operator setting. The scheduler immediately reports degraded coverage when the measured cycle exceeds that deadline; it never hides misses. Aggregate mailbox counters expose receipt, replacement, stale-drop, admission, deferral, and max-selection behavior without pixels, URLs, plates, or route data. A diagnostic-only helper also exposes per-camera received/replaced/drained/stale counts for pacing benchmarks; the operator surface remains aggregate.

7. Adaptive video analytics

```
SCOUT: periodic low-rate vehicle detection across every served camera
ENGAGE: increase sampling while a vehicle track is active
PURSUIT: prioritize cameras relevant to an active authorized investigation
CONFIRM: spend heavier OCR only on ambiguous high-quality evidence
DEGRADED: preserve priority work, reduce background service, expose coverage loss
```

An authorized designated-target arm publishes a bounded, latest-wins interest lease containing only evidence-backed observed camera IDs, a fixed expiry, and a replacement generation. The worker coordinator applies it on its owner thread; the scheduler gives only existing PURSUIT/CONFIRM work a bounded score bonus. It never receives the plate, route payload, or a target hash. Due SCOUT work still wins, and target interest cannot

manufacture CONFIRM; the runtime evidence gate remains authoritative. Expiry, replacement, empty evidence, and removed or unhealthy cameras clear the lease. Telemetry reports aggregate counts only. The optional hint is capped at 512 unique cameras; a larger statewide route receives no boost rather than delaying or rejecting the authorized trace.

The pipeline is intentionally cascaded:

- 1 Decode each active camera once.
- 2 Detect vehicles at a bounded scout rate.
- 3 Track between detector frames with class-aware, one-to-one center/scale motion association; ambiguous matches start a new track rather than merge.
- 1 Localize plates inside vehicle regions at original crop resolution.
- 2 Reject unsupported tiny or distorted crops before specialist OCR.
- 3 Sample independently per track using exact source time. Adjacent attempts inside the 200 ms evidence spacing are skipped; unsupported crops do not consume the bounded five-crop specialist budget. Scope, session and loop changes clear this metadata-only state.
- 1 Rank sharp, large and temporally diverse crops.
- 2 Run fast CCT OCR for scouting and selective Awiros verification.
- 3 Require each OCR path to satisfy exact support across at least three source-time-diverse frames, then require the two paths to agree exactly before automatic authority.
- 4 Preserve specialist-only and conflicting candidates for review; abstain from automatic alerts instead of correcting characters heuristically.

The bounded per-track crop budget preserves the current consensus gate but can change which later crops are available for ranking. It remains an optimization hypothesis until a labeled source-timestamped A/B replay proves unchanged exact track recall, alert precision and correct abstention; otherwise it is removed.

Measured warm-path MacBook inference places vehicle detection at roughly 95 to 128 ms per analyzed frame, plate detection at 20 to 25 ms per vehicle crop, CCT at 2.4 to 3.7 ms per selected crop, and Awiros at roughly 78 to 110 ms per selected crop across measured runs. The current same-build persistent-worker smoke loaded the float32 inference-exported RF-DETR/OIM/CCT bundle once in 4,243.628 ms, then resolved the same bundle from process cache in 0.005 ms; source-time-diverse detection-pipeline p50 measured 127.667 versus 111.078 ms. Four analyzed frames spanned 0.6 source seconds and could structurally supply four votes against the fixed three-vote minimum. A separate real-checkpoint lifecycle smoke loaded Awiros once in 4,903.225 ms, then resolved it in 0.001 ms and completed the warm prediction in 91.794 ms. These runs prove lifecycle reuse, source-time allocation and scheduling cost, not full-feed accuracy. The cascade exists because the detector dominates cost and most frames do not justify specialist OCR.

The retained 254-image external real-crop score measured Awiros at 80.315% exact single-crop recall. A 0.90 confidence floor still admitted 28 wrong strings among 231 accepted crops, so confidence is supporting telemetry only. The three-vote, 75%-support consensus accepted 8 of 26 repeated-identity groups: eight correct, zero wrong, and 18 abstentions. These are independent pre-cropped images with synthetic spacing, not full-frame or source-timestamped CCTV tracks; they do not establish production accuracy.

A separate selected lab replay ran nine real Indian plate crops through pinned OIM detection, Awiros verification, exact consensus, the watchlist engine and route backend. It produced three alerts and a complete three-point route. CoreML plate detection measured 9.132/21.442 ms p50/p95, specialist OCR 79.441/101.374 ms, and first replay frame through three calls plus persisted alert 277.058/319.921 ms. The identity was selected as known-resolvable, while cameras and times were synthetic. This proves the working inference-to-operator integration and its local latency, not accuracy or feed acceptance.

A fixed six-clip CC BY 4.0 Indian-road full-frame rehearsal tested the authority policy against every identity one reviewer could read in at least three of the published 1 FPS frames. The specialist-led baseline recovered four of

seven identities but automatically emitted one stable wrong identity and one duplicate track: 80% conditional precision. Requiring exact agreement between independently gated CCT and Awiros paths emitted two identities, both correct: 100% conditional precision, zero unsafe identities and zero duplicates, with 28.6% automatic recall. The five remaining identities were withheld from automatic alerts; one met the exact target-review candidate policy in that half-resolution report. The working local authority persists supported disagreements in a bounded, indexed review queue. After the designated registration is armed, only candidates at edit distance zero or one are shown, closest first. The operator UI locks its decision controls until a retained crop loads. A confirmation is append-only and projects the original unknown-identity record into the requested route with `human_confirmed_from_ambiguous_ocr` and `automatic_alert_withheld` labels. It never rewrites the OCR record and never enters watchlist matching. Browser acceptance proved confirmation adds one route point with zero alerts and that a later rejection removes it as an audited correction. Six clips and one reviewer do not establish population accuracy, route order or feed acceptance; they demonstrate why disagreement must never create an automatic alert.

The recovered fixed sample then isolated two production boundaries. A 3% plate crop margin first increased guarded recovery from one of seven to two. Keeping the probe's native 1920x1080 decode instead of the benchmark's earlier 960x498 decode raised automatic recovery to three of seven and exact target-review reachability to five of seven. Conditional precision stayed 100%, zero unsafe identities survived, and both negative clips abstained. A ten-source-second same-camera exact-identity debounce removed one duplicate caused by a measured seven-second tracking break; authority thresholds did not change. Selected specialist crops rose from 38 to 142, so the scheduler keeps Awiros bounded and just-in-time. This is licensed-media evidence, not supplied-feed accuracy or a broad population claim.

Face recognition is not in the mandatory path. It adds privacy, legal, accuracy and compute risk without improving the designated-vehicle test.

8. Sighting, watchlist and alert decision

A sighting is an evidence proposal, not a policing fact. Its canonical record includes:

- camera and confirmed source-system reference;
- stream session, loop and timing quality;
- source PTS/timebase, scenario time if known, and ingest time;
- raw OCR candidates, normalized exact string and consensus support;
- evidence reference and content digest;
- model artifact hashes, build identity and policy configuration.

The watchlist is append-only and versioned. Each entry has source, reason, validity, state and actor. Automatic correlation requires an exact normalized registration, an active validity window and an observed consensus. Fuzzy, ambiguous, expired or paused entries never create automatic alerts.

Alerts begin as proposed. An authorized officer can confirm, reject or leave them for review. Disputed OCR appears in a separate **Needs review** lane and is never styled as an alert. Every decision preserves actor, time, note and prior state. Duplicate frames or retries cannot create duplicate alerts for the same evidence.

Target-guided review is a separate authority path. It uses bounded edit distance only to rank ambiguous candidates for the registration the officer already entered; the distance is not a confidence score. The decision controls route membership only. A retained crop, active target, authorized case/purpose and append-only actor/time record are required. Direct API calls cannot promote an unretained or more-distant candidate, and no human review can create an automatic alert.

9. Journey reconstruction and operator behavior

The operator posts the designated registration in the request body; it never appears in a URL. The journey service returns ordered observed points with evidence, timing quality and uncertainty.

Route rendering follows five rules:

- Observed: exact supported sighting with source evidence.
- Human-confirmed: retained ambiguous evidence accepted by an authorized officer for this target; the original OCR record remains unknown and automatic alert authority stays withheld.
- Uncertain: supported sightings whose time intervals overlap or cross an untrusted boundary.
- Gap: camera coverage, source-system identity, GIS or feed health is insufficient.

■ Implausible transition: the minimum straight-line speed, even after applying timestamp uncertainty, exceeds a conservative road-vehicle bound. Preserve both observations, mark the route partial, and require source review.

The UI does not draw a confident line through missing evidence. It plots ordered route points on an offline coordinate surface only when registry latitude and longitude are provenance-backed; unresolved locations remain explicit. The officer can open the source reference and derivation for every material point, see alert state, and export a deterministic evidence report.

The live viewer is demand-driven. Selecting a registry record opens no stream. An operator may open at most two previews; closing a tile tears down its session. This bounds bandwidth, browser decode and gateway load while keeping the jury demonstration direct.

10. Data model and APIs

Primary records:

- CameraRegistryRecord and verified GIS overlay.
- StreamMeasurement and FeedSession.
- VersionedWatchlistEntry.
- PlateCandidate, TrackConsensus and Sighting.
- Alert and append-only AlertDecision.
- TargetReviewDecision, append-only and scoped to one sighting plus one designated registration.
- RoutePoint, Journey and EvidenceManifest.
- MetadataAudit, SearchAudit and ExportAudit.

Working API families:

```

POST /registry/catalogue/sync
POST /registry/cameras
POST /registry/cameras/bulk
POST /registry/measurements/import
GET /registry/cameras
GET /registry/cameras/{camera_id}
GET /registry/cameras.geojson
GET /registry/gaps
GET /registry/export.csv
GET /registry/audit
POST /journey/watchlists
POST /journey/sightings
GET /journey/review-candidates
POST /journey/designated-target/review-candidates
POST /journey/designated-target/review-candidates/{sighting_id}/decisions
POST /journey/tracks/consensus
POST /journey/search
GET /journey/alerts
POST /journey/alerts/{alert_id}/decisions

```

Mutations are idempotent or conflict-detecting. List endpoints are bounded. Secrets, raw URLs, cookies, frames and registration strings are excluded from health output and diagnostic logs.

11. Security and privacy architecture

Trust boundaries separate feed credentials, analytics, evidence, operator identity and optional cloud services.

- Deny sensitive operations by default; scope search, evidence, watchlist, export, registry and audit separately.
- Demo role tokens stay in browser session storage and same-origin headers. Production derives actor and department scope from SSO or mTLS claims.
- Inference credentials remain memory-only inside the feed adapter. The viewer never sends registry tokens to cross-origin media endpoints.
- Encrypt watchlists, selected evidence and backups in production; separate key administration from application roles.
- Segment operator, service, analytics and evidence networks.
- Keep raw video in departmental/regional storage where policy permits. Retain only the smallest evidence needed for review.
- Log access and decisions without copying frames or plate strings into logs.
- Optional NVIDIA NIM or Cloudflare services receive derived asynchronous data only and cannot modify the authoritative local route or alert path.
- A content digest proves evidence bytes are unchanged; it does not prove the model conclusion is correct.

Journey writes cross a bounded background queue. Decode and inference never wait on the registry API. The queue has hard item and byte limits, deterministic sighting deduplication, and separate queued, persisted, failed, duplicate and capacity-rejection counters. A controlled end-of-window barrier may wait for delivery; timeout or rejection makes the run visibly degraded. Only the registry's accepted response counts as persistence.

The live path is incremental rather than window-batched. One serialized warm runtime consumes supervisor jobs, keeps camera/session/loop track state isolated, copies only bounded plate crops, and releases each full RGB frame when the job completes or fails. The system selects scout, engage, pursuit, or confirm from privacy-safe track/crop state. Three verifier-supported source-time diverse candidates spanning at least 0.2 seconds admit specialist finalization; only a typed independently confirmed result becomes a sighting. Reconnect, loop, timebase change, unexplained PTS regression, inactivity, and stop close the old scope before new evidence can enter it.

The concrete adapter uses RF-DETR, OIM and CCT as one explicitly owned warm bundle. Awiros remains unloaded until CONFIRM. One control-plane constructor wires that adapter to the serialized runtime, bounded delivery driver and each camera's scope callback, preventing deployment code from accidentally choosing parallel model calls or bypassing lifecycle cleanup. Crop frame identity is stored inside the bounded crop evidence itself, avoiding session-long metadata growth.

12. Deployment architecture

12.1 Demonstration

The judged path is local-first:

```
government feed gateway
-> one inference host with a crash-isolated persistent model worker
-> local FastAPI registry/journey authority and SQLite WAL
-> local operator browser

support hosts
-> registry/health, analytics experiments, recovery and evidence metadata
```

Venue WAN loss must not break registry search, alert review, route construction or retained evidence. External services are optional accelerators only.

The final two-camera acceptance command uses two provenance-confirmed source systems and two distinct authorized full-frame recordings with genuine source-time offsets. One protected API and one persistent inference worker process both sources, then replay both. Acceptance requires cache transitions of cold/warm/warm/warm for the model bundle and verifier; track-level scout and specialist agreement on the target digest; at least three source-time-diverse winning evidence frames per source; exact configured locations; an independently evidenced alert for each route point; byte-exact regeneration of the deterministic report; unchanged source-backed inference on replay; and route, alert and watchlist persistence after an API restart. Any verifier failure, mismatched build, relabeled recording, aggregate-only count, missing alert, missing second observation, report tampering or uncertain source identity fails the run closed.

A separate deterministic incremental acceptance now drives two interleaved H.264/H.265 camera scopes through the real consensus gate, bounded background writer, watchlist matcher, alert store and route engine. It proves two early sightings before scope end, replay deduplication, full-frame payload release and restart persistence. Its analyzer and media are synthetic; it proves the runtime composition, not camera/model accuracy.

12.2 Production regional cell

Each cell owns feed sessions, adaptive inference, short evidence retention and local health. The state control plane owns the registry, global policy, searchable event index, route service and audit metadata.

Cells scale horizontally by camera assignment. A camera has one active decode owner and a warm standby lease. Control-plane partitions use source-system/region plus time. Metadata and selected evidence replicate asynchronously; raw video need not.

The current process refuses more than 64 assigned cameras. This is a bounded memory-safety envelope, not measured inference capacity. Within it, the queue keeps one latest scout per camera plus eight investigation slots. The regional benchmark selects the smaller operational assignment that meets latency and 30 percent headroom gates; a larger catalogue is explicitly sharded rather than accepted into one overloaded worker.

13. Statewide sizing and performance

Published target: approximately 80,000 cameras.

Input assumption	Statewide sensitivity
1 Mbit/s average	80 Gbit/s; 0.864 PB raw video/day
1.5 Mbit/s average	120 Gbit/s; 1.296 PB/day
2 Mbit/s average	160 Gbit/s; 1.728 PB/day
1 retained metadata event/camera/minute, 1 KB/event	115.2 million events/day; about 115 GB/day before indexes/replication

The reproducible Model 1 benchmark built 80,000 synthetic camera records through the audited SQLite path in 6.932 seconds. A bounded 30-row API page measured 76.333 ms p50 and 77.477 ms p95. Exact API lookup measured 1.662 ms p50 and 2.397 ms p95. Eight concurrent clients sustained 20.372 list requests per second with 438.579 ms p95 latency across 160 calls. The checkpointed database was 111,247,360 bytes and process peak RSS was 277,839,872 bytes, including fixture construction and the test runtime. These are metadata results, not video-capacity claims; the retained JSON artifact records the environment, method and limitations.

A separate synthetic decode-only rehearsal opened sixteen paced 1080p/25fps sessions, split evenly between H.264 and HEVC. Software and VideoToolbox each delivered all 3,200 expected frames. VideoToolbox reduced child CPU from 1.326 to 0.562 equivalent cores, about 58 percent, while sampled aggregate RSS remained about 1.8 GB. One forced missing-source attempt recovered on the next attempt after the two-second initial backoff. Generated test patterns are easier than CCTV and the run excludes network, inference and evidence sinks, so it selects a Mac demo decoder candidate; it does not establish a regional-cell capacity.

A real-model replay burst then used six licensed Indian-road clips across 30 and 50 logical aliases. Every frame was admitted under the corrected per-camera queue, but only 5/30 and 4/50 were analyzed before the two-second freshness gate; the remainder were dropped stale. Warm model-service p95 was 970.215 ms and 1,169.303 ms respectively, and peak process RSS was about 1.08 GB. This fails the 30 percent headroom gate and rules out one serialized CPU worker for synchronized 30/50-camera bursts. It does not establish sustained capacity, distinct-camera evidence, codec diversity or N+1.

A current sustained replay exercised the real coordinator callback path with one independent paced producer thread, exact source-PTS cadence, phase-spread first emissions, float32 RF-DETR inference export and measured model-service p50/p95/p99. Deadline-aware routine admission targets an eight-second sweep inside the ten-second scout deadline. Fifteen-, twenty-, and twenty-four-camera serialized Mac cells passed the fixed 1.30 gate at 2.200, 1.712, and 1.575 times admitted demand with zero loss or deadline miss. The 30/50 controls missed 22/51 deadlines. A host-contended run also missed at 24, so twenty remains the operational planning cap until 24 proves repeatable. The retained artifact uses six licensed recordings as aliases and is not government-feed, distinct-camera, accuracy, transport or N+1 evidence.

A balanced same-host fleet then ran two concurrent 15-camera cells with independent real model bundles and explicit four-thread Torch/ONNX CPU budgets. Both cells served every camera through at least two steady windows with no stale, capacity, deadline, or producer loss. Service reserve was 1.828 and 1.838 times admitted demand, proving the local 30-camera demo envelope. An independent ARM64 arminator CPU cell passed eight cameras for 30 seconds and three windows at 2.068 analyzed fps versus 1.200 admitted fps, with zero loss, deadline miss, or swap growth. The statewide topology still distributes inference cells beside regional feeds and sends asynchronous metadata/evidence to the state plane; these licensed aliases do not prove statewide scale, N+1, or government-feed capacity.

The production inference bill of materials remains intentionally unfrozen. One mixed-codec regional cell must measure:

- sustained decoded and analyzed frames per second;
- source-frame-to-alert p50, p95 and p99;
- stale drops, queue age, reconnect and degraded coverage;
- CPU, RAM, accelerator memory, network, disk I/O and power;
- exact track accuracy, correct abstention and false alerts;

- N+1 behavior with one worker unavailable.

Only the smallest cell that passes these gates with 30 percent inference headroom becomes the basis for statewide extrapolation.

14. Availability, disaster recovery and observability

- One active feed owner per camera; bounded reconnect near 2 seconds and capped near 30 seconds.
- Health state distinguishes catalogue presence, transport, decode, timing, inference and evidence sinks.
- Bounded queues expose queue age, stale drops and reduced coverage.
- Regional workers use leases and idempotent event writes to prevent duplicate ownership.
- Production registry, event, watchlist and audit stores require encrypted replicated backups and tested point-in-time recovery. The current prototype proves only a local full-backup boundary.
- The retained synthetic rehearsal restored all ten critical tables and the same three-point route, three alerts, operator report, schema digest and state digest in 23.086 ms for a 143,360-byte fixture. The append-only audit trigger blocked mutation and the state manifest detected deliberate alert tampering. This timing does not extrapolate to production volume.
- Control-plane loss leaves regional capture and bounded event buffering active; regional loss becomes a visible coverage gap.
- Recovery objectives are policy inputs until departments specify operational criticality and retention.

15. Cost-benefit decision

The cost driver is continuous pixel processing, not the metadata registry. The design reduces cost by:

- reusing existing cameras, VMS and departmental storage;
- decoding once per active camera;
- applying low-rate scout plus tracking instead of full analytics on every frame;
- invoking specialist OCR only on supported high-quality crops;
- opening previews only on operator request;
- moving metadata and selected evidence rather than duplicating all video statewide;
- keeping free NVIDIA/Cloudflare capabilities off the authoritative path unless a benchmark proves benefit.

Every optimization is removable. If it worsens exact track accuracy, alert precision, verified-result latency, recovery or total cost under the same labeled feed, it is rejected.

16. Department prerequisites

Each department or source-system owner must provide:

- accountable owner and escalation contact;
- confirmed camera-to-source-system mapping with evidence reference;
- VMS/NVR vendor, version, supported API/SDK and change window;
- feed protocol, codec, resolution, rate, time synchronization and network path;
- retention, playback and evidence-export capability;

- camera location and coordinate provenance;
- authentication, certificate, allowlist and rotation requirements;
- availability, bandwidth, outage and AMC constraints;
- privacy, legal basis, role scope, audit and retention policy;
- expected camera growth and planned replacement.

VAHAN, SARTHI, eGujCop/CCTNS, AFIS and NAFIS remain authorized adapter contracts until official credentials and schemas are supplied. The prototype must not claim live access.

17. Phased rollout

- 1 Sandbox proof: dynamic catalogue, two confirmed source systems, exact live ANPR, watchlist alert, route and evidence report.
- 2 Controlled PoC: one or two departments, measured regional cell, SSO/mTLS, retention policy, fault and restore rehearsal.
- 3 Regional expansion: connector certification, N+1 cells, event index, operational dashboards and support runbooks.
- 4 State control plane: registry, policy, cross-region route and audit federation.
- 5 Statewide scale: add cells only from measured capacity, accuracy and cost evidence.

18. Open risks and decisions

Risk or unknown	Current treatment	Closure evidence
Current gateway media edge returns HTTP 502	Keep local authority working; retry one bounded feed, do not fake video	Successful WHEP/HLS/RTSP session log and recording
Two source systems not exposed in current catalogue	Keep status inferred/unresolved; require provenance-confirmed mapping	Two confirmed registry records and simultaneous viewer proof
Exact venue bandwidth and cloud policy unknown	Local-first judged path; external services optional	Organizer/venue confirmation and WAN-loss rehearsal
No proven CUDA host in team inventory	Keep portable CPU path; rent/bring GPU only after same-feed benchmark	Measured cell passes accuracy, latency and headroom gates
Real track-level accuracy not yet measured on representative government plates	Use the external crop/consensus score only as bounded model evidence; abstain outside exact support	Labeled source-timestamped government windows and scored report
Exact deadline time and Model 5 wording conflict	Submit early; do not invent Model 5	Organizer clarification or saved receipt before cutoff

19. Submission acceptance run

The final rehearsal uses one immutable build manifest and produces:

- 1 Catalogue snapshot and registry gap report.
- 2 Browser recording of onboarding, viewer, health and GIS behavior.
- 3 Government-feed analytics report with vehicle/plate timestamps.
- 4 Representative watchlist version, automatic alert and officer decision.

- 5 Designated-vehicle route with timestamp/location evidence and visible gaps.
- 6 Mixed-codec, reconnect, discontinuity, bounded-queue and WAN-loss logs.
- 7 Exact accuracy, abstention, false-alert, latency and resource report.
- 8 HLD, standalone workflow diagram, presentation and videos generated from the same build identity.

The submission is not complete until every link works for a read-only viewer and every claim maps to retained evidence.

20. Submission traceability

Requirement	Design response	Evidence or remaining gate
R06 - secure statewide integration	Regional pixel plane plus state metadata/evidence control plane	HLD complete; production cell measurement remains
R10 - open, modular, vendor-neutral	Catalogue and connector adapters behind canonical APIs	Adapter contracts and replacement-safe boundary implemented
R13 - compulsory registry and GIS	Registry, safe onboarding, health, maintenance, audit, gap report and verified-only GIS	Working portal/API and 166-item sample gap report
R14 - direct viewing and analytics	Consume-only WHEP/HLS viewer plus direct inference adapter and ANPR metadata	Two confirmed live source systems remain
R20 - designated vehicle route	PTS-aware exact search with evidence-linked observed/uncertain points	Selected real-crop lab replay proves three inference-derived points; live multi-camera proof remains
R21 - watchlist alert	Versioned exact matcher, validity, deduplication and human decision	Three inference-derived lab alerts proven; continuous accepted-feed run remains
R23/R25 - protocol and source timing	RTSP/TCP inference, WHEP/HLS preview roles, PTS/timebase authority	HLS decode proven; WHEP media and RTSP/TCP remain
R30 - HLD	Components, ingest, analytics, watchlist, operator flow, scale, security, prerequisites	This document and standalone workflow PDF
R35 - approximately 80,000 cameras	Measured Model 1 benchmark, video sensitivity, repeated regional cells	Mixed-codec inference cell benchmark remains
R36 - security and privacy	Local authority, role boundaries, minimization, audit, tamper detection and no fuzzy auto-alert	Local full-backup restore proven; production SSO/mTLS, encrypted off-host backup and point-in-time recovery remain

21. References

- Official problem statement: <https://sentinel.gujarat.gov.in/problems>
- Official resources guide: <https://sentinel.gujarat.gov.in/resource>
- Official FAQ: <https://sentinel.gujarat.gov.in/faqs>