

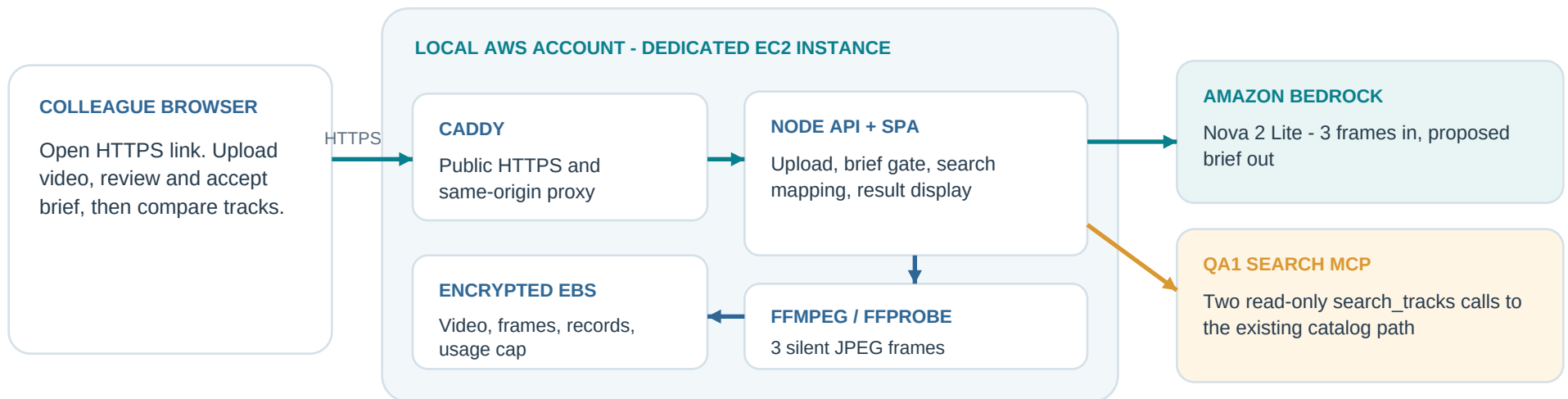
Video to search: what the link runs

One SPA, one temporary server, real Nova analysis, and read-only QA1 catalog results.

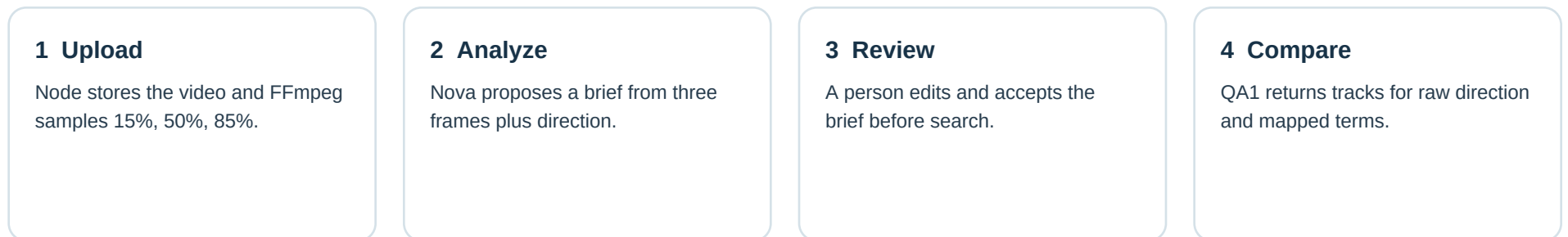
OPEN DEMO LINK

<https://video-demo.hgebrekidan.developer.apmmusic.com/>

RUNTIME TOPOLOGY



COLLEAGUE PATH



Boot-only artifact: a private S3 bundle supplied the app code. Browsers do not fetch the demo from S3.

Exact wiring and boundaries

This page describes the deployed POC. The production proposal begins on page 3.

HOP	ACTUAL DEMO WIRING	WHAT TO EXPECT
Browser to Caddy	Public HTTPS at video-demo.hgebrekidan.developer.apmmusic.com.	Anyone with the link can open it.
Caddy to Node	Reverse proxy to 127.0.0.1:8787 on the same EC2 instance.	The page uses relative /api calls.
Upload to FFmpeg	Up to 250 MB, 1-180 seconds; ffprobe plus frames at 15%, 50%, 85%.	No speech or audio detection.
Analyze to Nova	Bedrock Converse, US Nova 2 Lite inference profile; max 600 output tokens.	A proposed brief, not a track choice.
Accept to mapper	Reviewer edits keywords, scenes, summary; accepted version maps to search input.	Human approval gates search.
Search to QA1	JSON-RPC search_tracks twice: raw direction versus accepted keywords.	Read-only QA1 candidates; instrumental filter.
Records and budget	Encrypted EBS holds source, frames, JSON records and persistent usage counts.	Files remain until approved cleanup.
Private deployment	Private S3 bundle at boot; EC2 role grants only bundle read, Nova invoke, SSM.	No browser AWS keys or SSH port.

PUBLIC DEMO LIMITS

60 uploads, 40 Nova attempts, 120 comparisons total. Each comparison makes two QA1 calls. No login; usage counts are shared by all visitors.

WHAT THIS DOES NOT PROVE

No customer entitlements, full video understanding, audio analysis, preview/render, or Search Controller diverter path. Judge track relevance manually.

Proposed production: exploded AWS map

Lanes show ownership; arrows mark calls. Preview and final jobs branch after selection.

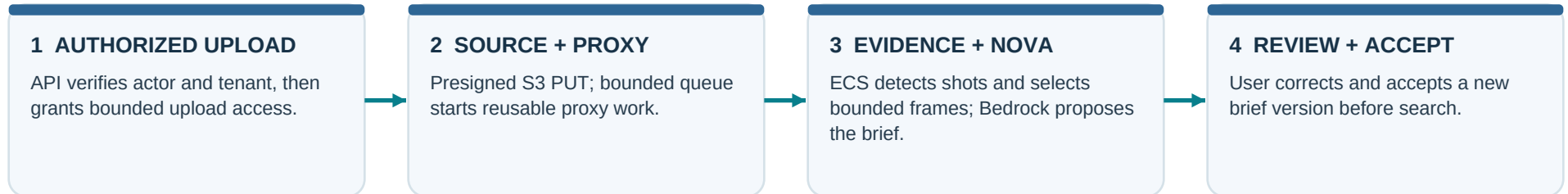


Shared controls across every lane: tenant policy, scoped IAM, encryption, audit, quotas, tracing and cost budgets.

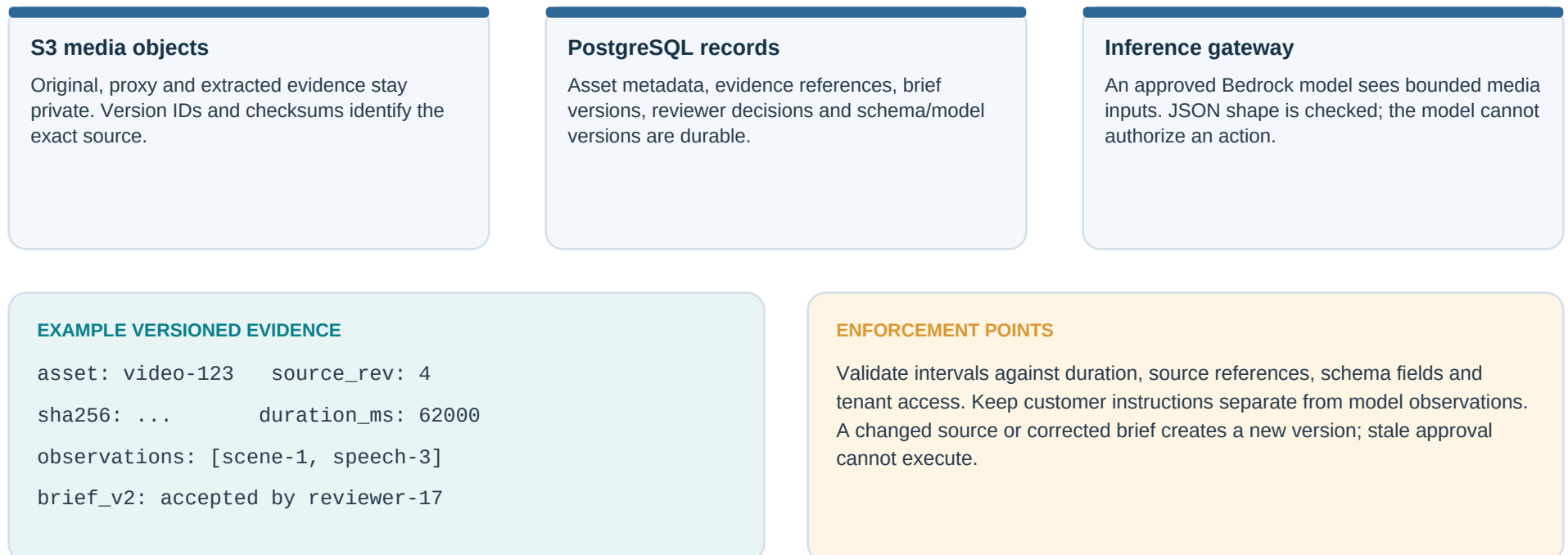
Intake to accepted brief

Sequential interpretation is an application workflow; MCP can expose context and search at reusable boundaries.

THE DEPENDENT PATH



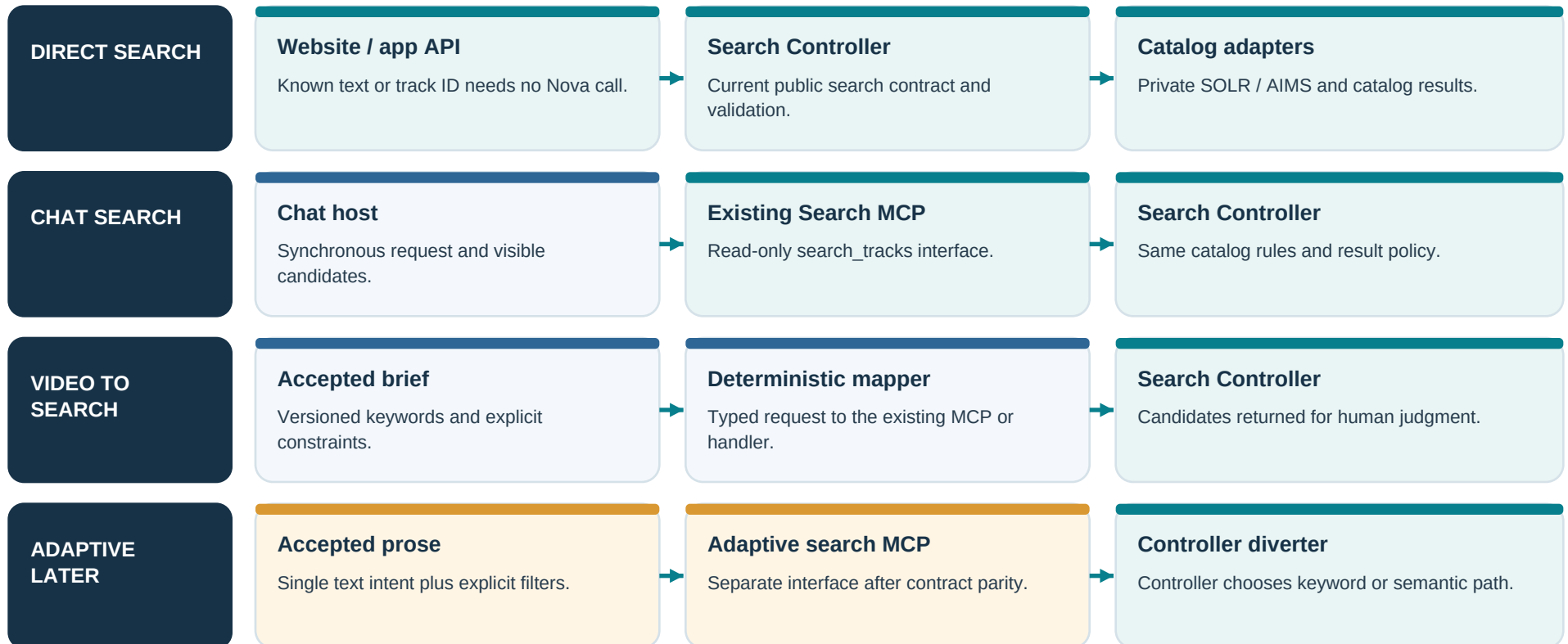
WHERE DATA AND DECISIONS LIVE



Shot evidence records boundaries, timestamps and selected frame IDs. Analysis can be polled; the accepted brief becomes typed search intent.

Search plane and policy owner

All candidate paths converge on Search Controller, which owns audience policy and private adapter decisions.



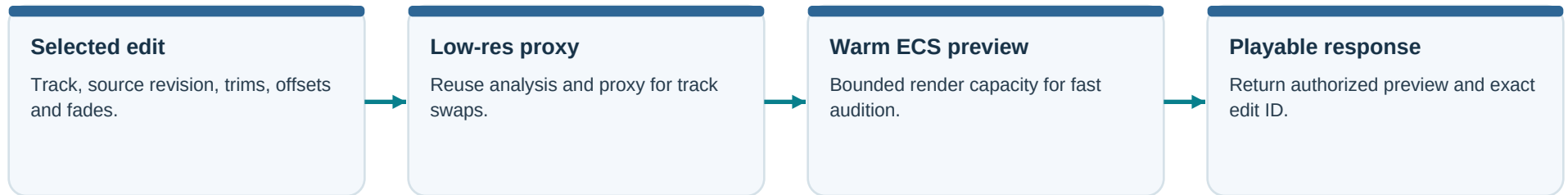
PRODUCTION GATE: IDENTITY, ENTITLEMENTS AND ROUTE EVIDENCE

Derive tenant and actor from verified application state; enforce catalog and asset rights at Search Controller and downstream services. Do not pass bearer tokens indiscriminately. The current Search MCP source uses anonymous catalog access, and its multi-term keyword request bypasses the Controller's single-text diverter. The adaptive tool must be tested separately with authenticated context, route logs, relevance and latency.

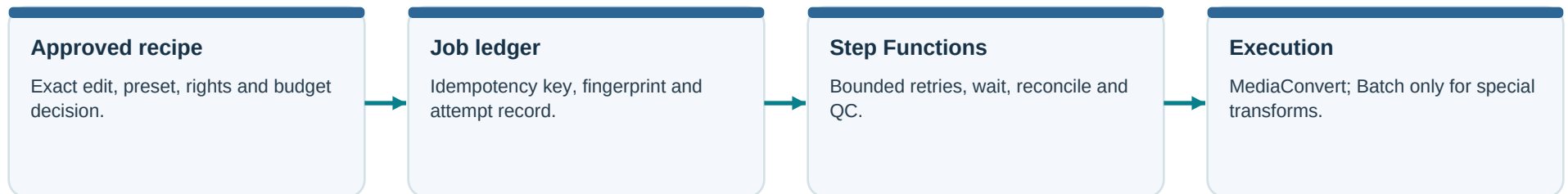
Audition, final render and delivery

The same selected edit can drive a quick preview and a durable final job; these are separate execution paths.

INTERACTIVE PREVIEW - RETURN IN THE CUSTOMER REQUEST



FINAL RENDER - ACCEPT JOB, THEN PROCESS DURABLY



OUTPUT AND AUTHORIZED DELIVERY

Technical QC produces a result manifest: output references and hashes, source/track/edit versions, provider job ID and cue-sheet facts. Private S3 stores the final. The application checks rights before issuing short-lived CloudFront signed access. Licensed source-asset packages, if offered, are a separate operation and receipt.

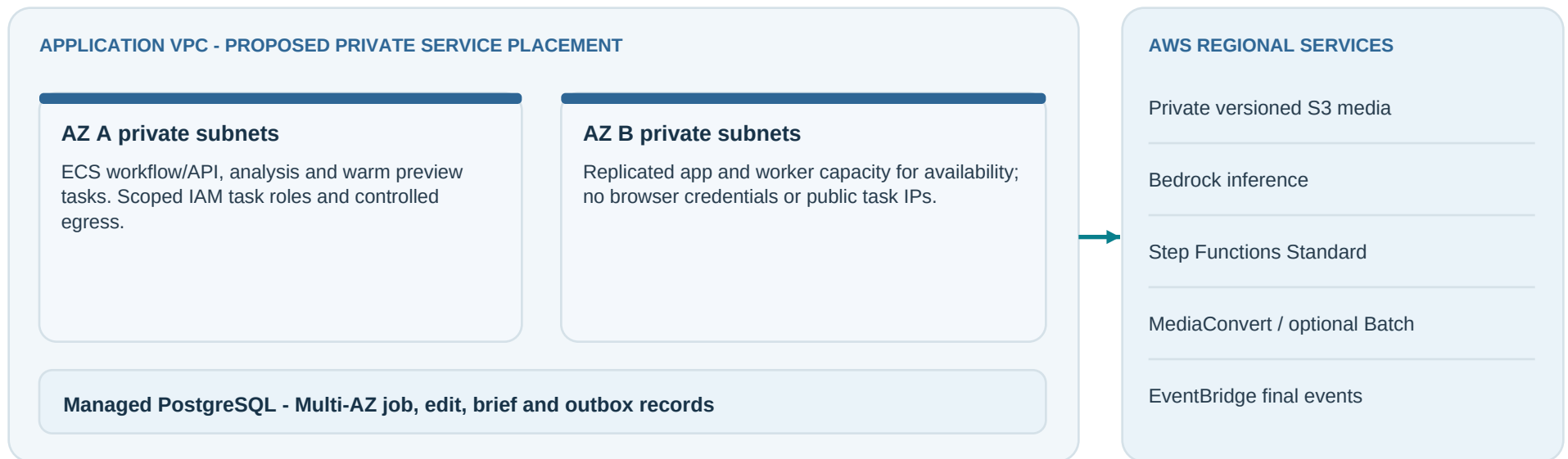
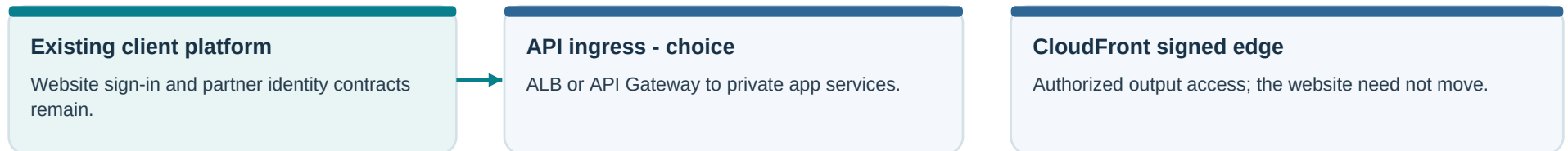
AUTHORITATIVE STATUS AND NOTIFICATION

PostgreSQL holds the job state across browser reconnects. A transactional outbox publishes final-render transitions to EventBridge; the existing Notification API pushes website updates over WSS. Clients fetch the authoritative job and result references through an API. Duplicate or out-of-order events cannot move state backward.

Candidate AWS deployment topology

A hosting and network view around existing APM services; account and ingress decisions remain open.

CUSTOMER EDGE AND API ENTRY



CROSS-SERVICE BOUNDARY AND OPEN PLACEMENT DECISIONS

Search MCP, Search Controller and Notification API retain their existing deployments; the new app calls their authenticated contracts. Use private S3 with CloudFront origin access control for delivered media. Review VPC endpoints, outbound egress, Region, account/cell strategy, final ingress choice and failure isolation with platform owners before implementation.

AWS foundation and trust boundaries

A proposed deployment shape for the full platform; exact account, VPC and tenancy layout require design review.

Edge and identity

Keep existing customer sign-in and partner identity contracts. Use CloudFront for static/authorized media delivery and an application API front door. The platform owner selects ALB/API ingress and any WAF policy.

Private compute

Run workflow, analysis and warm preview services on ECS across Availability Zones. Scope task roles by service. Keep AWS credentials and model access on the server side; use dedicated media execution roles.

Stores and encryption

Private versioned S3 holds originals, proxies, evidence and outputs. Managed PostgreSQL holds compact records and outbox. Encrypt at rest with managed keys; treat indexes as rebuildable projections.

Tenancy and rights

Derive tenant, actor and allowed operation from authenticated state. Apply entitlement checks at search, preview, final submission and delivery. Use tenant-aware keys and records; introduce tenant cells if policy or scale requires.

Recovery and audit

Use immutable input versions, idempotency keys, bounded retries and provider reconciliation. Emit trace IDs and structured audit data to CloudWatch/CloudTrail. Preserve accepted records through rollback.

Engineering delivery

Keep GitHub source and Azure Pipelines. Use OIDC to obtain short-lived AWS deployment roles. Separate nonproduction and production environments, use flags, measure cost and reliability, and gate releases on parity.

CONDITIONAL CAPABILITIES

A vector/OpenSearch projection, managed MCP gateway, GPU pool, or separate tenant cell enters only after workload, relevance, compliance or ownership evidence supports it. Define media retention, model/Region approval, egress rules, availability objectives and cost ceiling before choosing final network topology.

What exists, what is proposed, what to prove

This roadmap keeps the temporary demo from being mistaken for the complete AWS production platform.

STATUS	CAPABILITY AND OWNER	NEXT PROOF
EXISTING	Search Controller, Search MCP and website search	Map current contracts and baseline.
EXISTING	Notification API and website WSS features	Agree final-event contract only.
POC VERIFIED	Open demo: 3 frames, Nova brief, QA1 search	Judge real clips and customer relevance.
PROPOSED	Versioned evidence, brief, identity and policy	Validate tenant/rights and correction flow.
PROPOSED	Warm preview, durable render and signed delivery	Test edit fidelity, recovery and QC.
OPTIONAL	Adaptive MCP, index, gateway or tenant cells	Add only after parity or scale evidence.

STAGED ACCEPTANCE GATES

1 BASELINE Corpus, contract mapping, owners and thresholds.	2 VIDEO SEARCH Relevance, edits, latency, cost and rights.	3 AUDITION Correct version, fast swaps, cache access.	4 FINAL Idempotency, QC, recovery, final-only events.	5 AUTOMATE Policy limits, BI reconciliation, partner receipts.
---	--	---	---	--

Sources: APM platform proposal v2 (19 Oct 2026); local repository contract review (21 Oct 2026).
Reference: MCP architecture | AWS Step Functions | MediaConvert integration | CloudFront private delivery.