Edge Cache & Token Design Checklist

Before you ship a streaming catalog: design the cache key so segments stay shared, place and validate the token without keying on it, set an invalidation strategy, and prewarm for live. Engineering guidance — confirm CDN defaults live, they change.

1 · DESIGN THE CACHE KEY (path only)

- ☐ **Key on hostname + path, nothing else.** The default key on every major CDN is domain + URL path; that is what lets one cached segment serve everyone. Two viewers of seg_00042 must produce the same key.
- ☐ **Keep per-viewer values OUT of the key:** token / session id, auth cookie, User-Agent, analytics params. Each one allowed in forks every segment into a per-viewer copy.
- ☐ **Use Vary with discipline.** Vary only on low-cardinality headers you truly serve differently; never Vary: User-Agent or Vary: *. Your key fields: _____

2 · PLACE & VALIDATE THE TOKEN (without keying on it)

- ☐ **Prefer a signed cookie** (or signed header) for segments — it rides every request automatically and is excluded from the cache key by default. CloudFront recommends signed cookies for multi-file content.
- ☐ **If you must use a query-string token, exclude its params from the key.** CloudFront omits query strings by default; Akamai uses a Cache Key Query Parameters behaviour. The edge still validates it.
- ☐ **Sign the manifest, not every segment.** One short-lived, path-scoped permit covers the segments beneath it. Token placement chosen: _____

3 · INVALIDATION STRATEGY

- ☐ **TTL does the routine work:** long max-age on immutable segments, short on the live manifest. Set these two right and you rarely purge.
- ☐ **Version by new path for ordinary changes** — a new key fills cleanly, no purge, no half-updated edge. Purge is for removal and emergencies.
- ☐ **For removal/takedown, use surrogate keys (cache tags):** tag a title's renditions+segments with its content id, purge in one call. Use soft purge to refresh without a miss storm.

4 · PREWARM FOR LIVE & MULTI-CDN PARITY

- ☐ **Prewarm the edge before a premiere.** Push the opening segments + manifest into edge caches so the simultaneous start lands on hits, not a thundering herd against the origin (keep an origin shield too).
- ☐ **Token parity across CDNs.** Each CDN signs with its own scheme; a steered viewer can be locked out at the second gate. Plan a Common Access Token (SVTA/CTA-WAVE) if multi-CDN is on the roadmap.
- ☐ **Token auth is access control, not encryption** — add DRM for premium catalogs. Shield on? ___ token parity? ___ prewarm plan? ___

THE ONE SANITY CHECK BEFORE YOU SHIP

The cache key identifies the content; the token identifies the viewer; never put the second in the first. Worked: 100,000 concurrent viewers pull a fresh 4-second segment, so the edge sees ~25,000 segment requests/second. With a correct path-only key, each segment is fetched from origin ONCE and served to all — offload ~99.99%, origin barely notices. Put a per-viewer token in the key and every viewer's request is a unique key — ~100,000 origin fetches per segment, offload ~0%, and the origin is asked to serve 25,000 req/s of full-bitrate video it cannot, so it falls over during the premiere you cannot re-run. Field reports describe a 98% hit ratio dropping below 40% within minutes. The fix is not a CDN setting to find under pressure; it is keying on the path and keeping the token out. Figures illustrative; confirm CDN defaults and names live, they change.



GO DEEPER · THE 7-VOLUME HANDBOOK

Fora Soft Video Engineering Handbook

Seven volumes, 9,000+ pages: encoding, streaming, audio and AI for video products — from the pixel to production.

By Nikolay Sapunov, CEO at Fora Soft · Kindle & paperback on Amazon

Browse the handbook on Amazon →

This checklist keeps the cache key on the path, the token out of it and the edge warm before a premiere. Our seven-volume handbook covers the delivery stack this runs on; start with Volume 3 for CDNs, caching and delivery cost, then Volume 2 for HLS and DASH segments, and Volume 1 for encoding.

WHAT YOU WILL FIND INSIDE

- CDN caching and cache keys (Vol. 3)
- HLS and DASH segments (Vol. 2)
- Delivery cost and origin offload (Vol. 3)
- Encoding ladders (Vol. 1)

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI
Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF
Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery
Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing
Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs
Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video
Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Origin load spikes every time a live premiere starts?

We will fix your cache keys and token placement, plan invalidation and prewarm the edge across your CDNs. 250+ video products shipped since 2005.

Book a 30-min call → **WhatsApp →** **Email us →**