

Live vs VOD vs Near-Live — Decision Card

One page. Three categories. The deadline-driven trade-offs and the protocol families that fit each.

VOD

Content exists before play

Latency: not measured
Protocol: HLS / DASH / CMAF
Buffer: 20-60 s
Cache hit: ~99%
Optimise: cost, quality
Examples: Netflix, YouTube, lecture archives

Failure mode:
 egress bill, codec reach

Near-live

Live with deliberate delay

Latency: 5-90 s (deliberate)
Protocol: HLS / DASH / CMAF
Buffer: 6-18 s
Cache hit: ~95%
Optimise: predictable delay
Examples: sports betting, live shopping, linear TV

Failure mode:
 drift, moderator overload

Live

Every stage races a clock

Latency: 0.2-30 s (target)
Protocol: LL-HLS / WebRTC / MoQ
Buffer: 0.05-6 s
Cache hit: 5-85%
Optimise: latency, scale
Examples: Twitch, sports, telemedicine, auctions

Failure mode:
 stalls during peak moments

Three questions to pick a category

- Q1.** Does the content exist before the viewer presses play? **Yes -> VOD.** Otherwise continue to Q2.
- Q2.** Is sub-second interactivity required (bidding, conferencing, telemedicine)? **Yes -> Live, sub-second (WebRTC / MoQ).** Otherwise continue to Q3.
- Q3.** Is a deliberate 5-90 s delay acceptable (betting, moderation, broadcast graphics)? **Yes -> Near-live (HLS / DASH / CMAF with 6-18 s buffer).** Otherwise: **Standard / LL live (HLS, DASH, LL-HLS, LL-DASH).**

Rule of thumb

Tightening latency by 10x roughly doubles unit cost. Most 'we need low latency' requests are actually near-live in disguise. Pick the loosest deadline your product can live with — and design the architecture around it.

Primary standards (per Fora Soft Learn editorial source-hierarchy)

HLS: IETF RFC 8216 (Aug 2017). LL-HLS: Apple HLS Authoring Specification, rev 2025-09. DASH: ISO/IEC 23009-1:2022. CMAF: ISO/IEC 23000-19:2024. WebRTC: W3C Candidate Recommendation + IETF RFC 8825-8866. QUIC: IETF RFC 9000 (May 2021).