Throughput-Based ABR Tuning Sheet

Four knobs, three defaults, five failure modes — the working reference for shipping throughput-based ABR.

1. The four knobs

Window size K	How many recent samples the estimator uses. K=3 reacts fast; K=8 is smooth. Default K=5 for VoD, K=3 for live.
Safety factor	Divisor on the smoothed estimate. 1.0 (hls.js default), 1.25 (most common), 1.5 (mobile, satellite). Above 2.0 wastes bandwidth.
EWMA α	Forgetting rate, 0–1. Low α (0.1) = slow & steady; high α (0.5) = fast & reactive. Run two in parallel: slow for climb, fast for drop.
Drop hysteresis	How many bad samples before switching down. 1 = reactive; 2 = common; 3 = conservative. Always pair with a buffer-floor rule.

2. Deployment defaults

Deployment	K	Safety	EWMA α (slow/fast)	Drop hyst.
VoD, residential fibre	5	1.25	0.10 / 0.50	2
Live, mixed networks	3	1.4	0.15 / 0.50	1
Mobile-first OTT	4	1.5	0.10 / 0.40	1 + floor

3. Where throughput-based wins

- Stable networks (fibre, office Wi-Fi, satellite). Estimator converges in 2–3 segments.
- Short content (≤ 30 s). Faster ramp than buffer-based, which is still climbing when the video ends.
- Low-resource players (older smart TVs, set-top boxes). One division per segment fits anywhere.

4. Five failure modes & fixes

Bursty Wi-Fi	Estimator oscillates with throughput swings. Fix: raise EWMA α toward 0.1 (slower forgetting) or move to buffer-based for known-bursty audiences.
Mobile handoff	5× jump or drop in one segment is invisible to the window for 2–3 segments. Fix: pair with a hard buffer floor (drop to bottom rung if buffer < 2 s).
CMAF chunked transfer	Partial segments arrive at the rung's bitrate, not the network's rate. Player never climbs. Fix: measure throughput only on inter-chunk idle gaps.
Shared bottleneck	Multiple players on one router each think they own the pipe. Fix: use PANDA-style probing, or accept the noise and tune for stability.
First segment	Zero history; TCP slow-start inflates the first sample. Fix: persist last-session throughput; use a conservative middle rung if no history.

5. Pitfalls to avoid

- Trusting segment 1: discard or downweight it (TCP slow-start inflates it 2–3×).
- Tracking only rebuffer ratio: a throughput player can oscillate without rebuffering. Track switches/min.
- Ignoring the buffer entirely: add a buffer-floor rule even if you stay throughput-based.
- Hard-coded ceilings (e.g. 4 Mbps for mobile): break on 5G. Use network-type signal instead.
- One estimator for both directions: use two — fast for drop, slow for climb.

6. Standards anchor

- RFC 8216 §4.3.4.2 — multi-variant playlist the algorithm consumes.
- Apple HLS Authoring Specification rev. 2025-09 §4.7.5 — initial rung selection.
- ISO/IEC 23009-1:2022 — DASH MPD; DASH-IF guidelines for adaptation profile.
- ISO/IEC 23000-19:2024 — CMAF chunked transfer (breaks naive estimators).