

AI Deployment Topology Cheat Sheet

Latency budget × Deployment topology × Real-time vs batch — for every video-AI feature.

1. Latency budgets by user-experience class

User-experience class	Latency budget
Live captions (partial)	300 ms
Live captions (final)	1.0 s
Live translation	800 ms
AI agent in call	200–500 ms
Background blur per frame	33 ms
In-call action items	3–10 s
Meeting summary	30 s – 5 min
Highlight reel	1–10 min
Archive search index	hours / nightly

2. Topology layers — round-trip + cost-per-minute envelope

Layer	Round-trip	Cost / min / stream	Typical feature
On-Device	0 ms	\$0 / min	Background blur, on-device blur, noise
Edge	5–30 ms	\$0.0003–0.003 / min	Live captions, translation, in-call agent
In-Region Cloud	30–100 ms	\$0.001–0.05 / min	Meeting summary, generative video
Cross-Region Cloud	100–300 ms	\$0.001–0.05 / min	Regulatory or model-availability only

3. Five-step decision tree

1. What is the user-experience latency budget?
2. Max tolerable round-trip = budget – inference window.
3. Which topology layer fits the model size?
4. Cost-per-minute envelope at expected concurrency?
5. Streaming partial answers, or batched single answer?

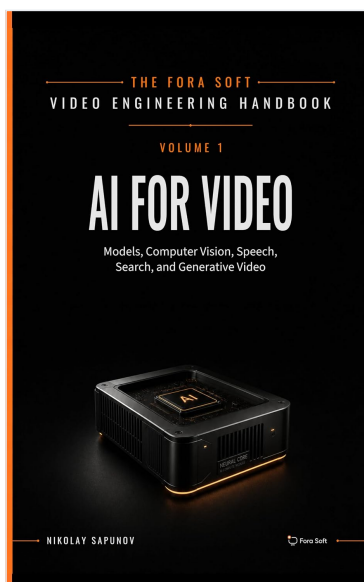
4. Three recurring mistakes

- On-device-as-cloud:** blur in cloud GPU when on-device costs \$0.
- Live-summary:** summarising mid-call when batched post-call is better.
- Cross-region default:** us-east-1 for EU users — fatal for real-time.

Companion to: 'Latency, Deployment Topology, and the Real-Time-vs-Batch Decision'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/latency-deployment-real-time-edge-cloud

Fora Soft · www.forasoft.com



GO DEEPER · VOLUME 6 OF 7

AI for Video: A Complete Guide to Intelligence in Video Products

From Models and Inference to Computer Vision, Speech, Search, and Generative Video

By Nikolay Sapunov, CEO at Fora Soft · 1,444 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6TKPZ1

This sheet maps latency budgets to on-device, edge and cloud layers for every video-AI feature. Volume 6 covers the decisions behind it: the AI feature lifecycle, deployment topologies, model artifacts and the ML basics you need to place computer vision, speech and search features.

WHAT YOU WILL FIND INSIDE

- Deployment topologies: device, edge, cloud
- Model artifacts and what they cost to run
- Computer vision and speech features in video
- Search and RAG over a video archive

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

Unsure where each AI feature in your video product should run?

We will place each feature on device, edge or cloud against its latency budget and estimate cost per minute. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)