The Edge + Cloud Reference Architecture - Build Pack

A one-page blueprint for a hybrid AI surveillance system: the six layers, the ONVIF standards spine, the budget for a 50-camera site, and the three configurations. Representative 2026 figures.

A. The six layers (left to right, over the LAN/WAN line)

- Capture: IP / AI cameras; an edge-AI NPU detects on-device in ~25-100 ms (ONVIF Profile S/T).
- Network and ingest: PoE switches on a wired LAN; the VMS pulls each stream over RTSP/RTP.
- Recording stays local: continuous to disk plus a 24-72 h ring buffer (ONVIF Profile G).
- Analytics in three tiers - camera NPU, on-prem GPU (16-40 streams), cloud; cloud also does fleet, re-ID, search, cold storage.

B. The standards spine (what keeps it vendor-neutral)

- ONVIF Profile S/T carries the live stream into the VMS (T adds H.265 and metadata streaming).
- ONVIF Profile G standardizes recording, search, and playback; Profile M carries analytics metadata.
- A Profile M consumer can be a camera, a server, or a cloud service - one metadata language across the line.
- RTSP (RFC 7826) and RTP (RFC 3550) carry transport; IEC 62676 frames the system. Conformance is a baseline, not every feature.

C. The budget - a 50-camera site

- Storage: ~32 TB local for 30-day continuous H.265 (2 Mbps = 21.6 GB/cam/day, x 50 cams x 30 days).
- Bandwidth: ~100 Mbps streaming all video vs ~5 Mbps hybrid - a ~95% upload cut.
- Analytics $/cam/month: ~$3 on camera, ~$8 on-prem GPU, ~$45 rented cloud GPU, ~$4,350 per-minute API.
- The rule: heavy + continuous work (recording, constant detection) local; occasional + heavy (search, cross-camera) cloud.

D. Three configurations, and the design rule

- Single-site: one VMS + GPU appliance does ingest, recording, and edge analytics together.
- Multi-site: per-site recording (so an outage at one site loses nothing) federated through the cloud.
- Remote / cellular: local storage is the primary record; store-and-forward sends key events when the link is up.
- Design top-down from timing, privacy, and link constraints - recording stays on the local network on every path.

Privacy + retention wraps every layer: keep recognizable video local and send the cloud only metadata and short clips - GDPR data minimisation (Art. 5(1)(c)) expressed as an architecture. Biometrics (face, plate) are a legal gate, not just a feature - the EU AI Act (Reg. (EU) 2024/1689; Annex III high-risk duties due 2 Aug 2026, a proposed Digital Omnibus may postpone to Dec 2027) and Illinois BIPA (740 ILCS 14) - keep that processing on hardware you control and get privacy/legal sign-off before shipping. Engineering guidance, not legal advice; confirm specifics with qualified counsel. Figures are representative for 2026 and scale with retention, resolution, frame rate, and model size. Sources: ONVIF Profiles S/T/G/M; IETF RFC 7826 and RFC 3550; IEC 62676; GDPR; EU AI Act; Illinois BIPA; NVIDIA Metropolis; AWS IoT Greengrass; Eagle Eye Networks; Verkada.