On-Prem AI Server - Sizing Checklist

Size an edge server by streams of your real model, not by a spec-sheet TOPS number. Decode budget first, inference second, headroom always.

A. Size by streams, not by TOPS

- [ ] Count the cameras and their resolution - 4K loads a GPU far harder than 1080p.
- [ ] Decode first: confirm the GPU's hardware decoder count and streams per decoder.
- [ ] Inference next: get streams-per-GPU for YOUR model, at YOUR resolution and fps.
- [ ] Assume ~16-40 light-detector streams per GPU; heavy models (face, re-ID) far fewer.

B. Pick the hardware

- [ ] Match the GPU class to the workload (mainstream inference GPU, or a small edge module).
- [ ] Confirm GPUs per server and that the room has the power and cooling for them.
- [ ] Decide build-your-own (NVIDIA-Certified) vs a turnkey appliance - control vs speed.
- [ ] Confirm the analytics software supports your GPUs and runs fully on-premise.

C. Plan for failure and growth

- [ ] Treat one server as a single point of failure - plan a second box or a failover path.
- [ ] Confirm you can add GPUs or servers as the camera count grows.
- [ ] Confirm the VMS / analytics federate across multiple servers as one system.
- [ ] Weigh the one-time server cost against the recurring cloud egress it replaces.

D. Residency and compliance

- [ ] Confirm whether the video must stay in-building or in-country (GDPR Chapter V).
- [ ] For biometric workloads, confirm the legal basis FIRST (EU AI Act, GDPR Art. 9, BIPA).
- [ ] Confirm recognizable footage never leaves the boundary if a residency rule requires it.
- [ ] Document the retention period and who can access the on-prem analytics output.

Remember: a GPU's fixed hardware decoders fill up before its AI compute does, so a server sized on TOPS alone will carry fewer cameras than the spec suggests. On-prem wins when the model is heavy or cross-camera, the data must stay local, or egress costs more than the box. Engineering guidance, not legal advice. Sources: NVIDIA L4 / DeepStream / Jetson documentation; GDPR Reg. (EU) 2016/679 Ch. V & Art. 5; EU AI Act Reg. (EU) 2024/1689; Illinois BIPA 740 ILCS 14; ONVIF Profile M.