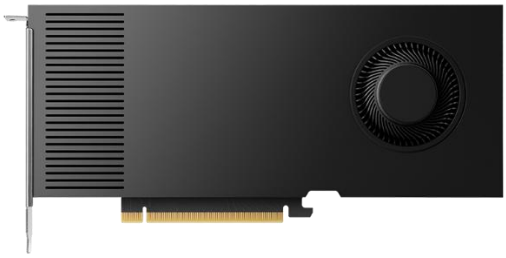
1. Typical Edge AI System – Hardware Configuration

🔹 A. Light Commercial Edge (Dental clinic / Surrogacy clinic / Small office)



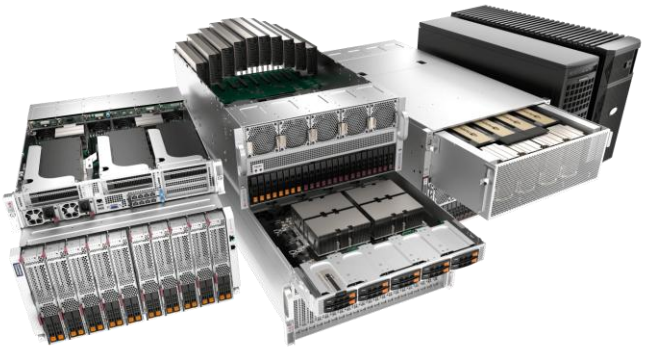
✅ Hardware Specs

Component	Typical Spec
CPU	Intel i7 / Xeon E-series OR AMD Ryzen 9
RAM	64GB DDR5 (minimum 32GB)

Component	Typical Spec
GPU	NVIDIA RTX 4000 Ada / RTX 4080
Storage	2TB NVMe SSD (Gen4)
Backup	4TB SATA SSD or NAS
Network	Dual 2.5G Ethernet
Power	750–1000W Platinum PSU
Chassis	Tower or 1U rackmount
UPS	1500VA enterprise UPS

💰 Budget range: **$4,000 – $8,000**

🔹 B. Enterprise Edge (Multi-site clinic / Travel agency HQ / Law firm)



4

✅ Hardware Specs

Component	Typical Spec
CPU	Dual Xeon Silver/Gold
RAM	128–256GB ECC
GPU	NVIDIA L40S / A100 40GB
Storage	4TB NVMe RAID 1
Backup	NAS or remote sync
Network	Dual 10G Ethernet
Rack	12U–24U rack cabinet
Redundancy	Dual power supply

💰 Budget range: **$25,000 – $80,000**

This is what impresses enterprise clients when you say:

> “We operate local AI inference nodes with GPU acceleration.”

---

2. Typical Edge System – Software Stack

Now the important part — what actually runs on it.

🔹 Operating System

- Ubuntu Server 22.04 LTS

- Rocky Linux (enterprise alternative)
- Windows Server (if Microsoft stack needed)

For your AI deployment strategy, I recommend:

> **Linux-based stack** (more stable + GPU optimized)

🔹 Core AI Software Layer

Layer	Typical Software
Containerization	Docker
Orchestration	Docker Compose / Kubernetes (light)
GPU Runtime	NVIDIA CUDA + cuDNN
AI Framework	PyTorch / TensorFlow
LLM Runtime	Ollama / vLLM / Triton Inference Server
Vector DB	FAISS / Milvus / Weaviate
RAG Layer	LangChain / LlamaIndex
API Layer	FastAPI / Node.js
Monitoring	Prometheus + Grafana

🔹 Voice AI (For Clinics / Phone AI)

If the clinic wants:
“AI answering phone calls locally”

Software stack would include:

- SIP server (Asterisk or FreePBX)
- SIP trunk provider integration
- Whisper (speech-to-text)
- Local LLM inference
- TTS (Coqui TTS or ElevenLabs cloud hybrid)
- Call routing logic

Now you truly have:

> A real AI receptionist running inside the clinic.

That is impressive.

---

3. Optional “Marketing Layer” (For Impressing Clients)

You can brand it as:

PanAuro Edge AI Node™

Include:

- Branded rack label
- Dashboard display on iPad
- Real-time GPU utilization monitor
- “AI Status Online” panel

Clients love visible infrastructure.

---

4. Edge vs Cloud Strategy (For Your Business Model)

For PanAuro:

Edge	Cloud (PES AI Factory)
Low latency	Heavy training
Sensitive data stays local	Large batch compute
Marketing credibility	Scalable
Small inference jobs	Model updates

This hybrid model is powerful for your:

- Travel industry pitch (GTT dinner)
- Clinics
- Professional firms

---

5. Minimum Edge Kit You Can Standardize (Recommended SKU)

For scaling your channel partners:

📦 PanAuro Edge Starter Kit

- RTX 4000 Ada
- 64GB RAM
- 2TB NVMe
- Ubuntu + Docker preinstalled
- Preconfigured AI agent stack
- Remote management enabled

You can ship this as a product.

---

6. Important Strategic Note (For You Personally)

Keith — given your GPU supply channel and PES infrastructure:

You do NOT want to oversell hardware.
You want:

> “Edge for presence, Cloud for power.”

Edge = credibility
PES AI Factory = revenue engine

That keeps capex manageable while building a strong AI narrative.