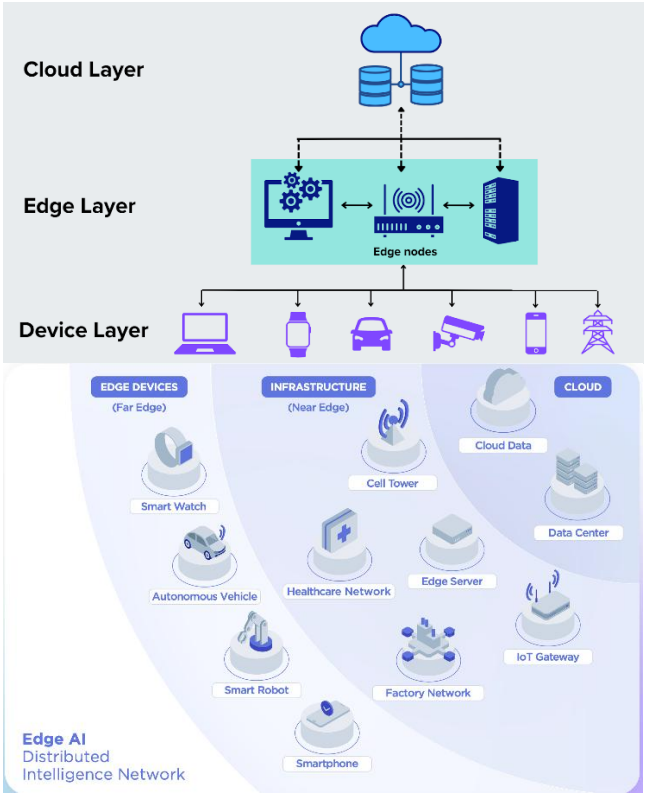
📘 PanAuro Edge AI Node™

Standardized Deployment Manual

Version 1.0 – Enterprise Deployment Framework

---

1. Purpose & Architecture Overview



4

1.1 Objective

The **PanAuro Edge AI Node™** provides:

- Low-latency AI inference
- On-premise data privacy
- Hybrid cloud synchronization with PES AI Factory
- Enterprise AI automation (voice, RAG, workflow)

It is designed for:

- Clinics (dental, surrogacy, medical)
- Travel agencies
- Law firms
- Accounting firms
- Logistics offices
- Enterprise branch offices

1.2 Hybrid Architecture Model

Edge Layer (On-site)

- Real-time inference
- Voice AI
- Document processing
- Sensitive data handling

Cloud Layer (PES AI Factory)

- Heavy model training
- Model updates
- Batch analytics

- Fleet management

---

2. Standard Hardware Configuration

2.1 PanAuro Edge Starter (SME Deployment)

Compute

- CPU: Intel i7 / Xeon E
- RAM: 64GB DDR5
- GPU: NVIDIA RTX 4000 Ada
- Storage: 2TB NVMe (RAID optional)

Networking

- Dual 2.5G Ethernet
- Firewall appliance (optional)

Power

- 1500VA UPS
- Surge protection

Form Factor

- Tower or 1U rackmount

2.2 Enterprise Edge Node

- Dual Xeon
- 128–256GB ECC RAM
- NVIDIA L40S / A100
- 4TB NVMe RAID 1
- 10G networking
- Redundant PSU

---

3. Software Stack Standardization

3.1 Base Operating System

- Ubuntu Server 22.04 LTS
- Hardened configuration
- SSH key-only login
- UFW firewall enabled

3.2 Core Platform Layer

Layer	Software
Container Runtime	Docker
GPU Runtime	NVIDIA CUDA Toolkit
AI Framework	PyTorch
Inference Engine	vLLM / Triton
Vector Database	Milvus / FAISS
RAG Framework	LangChain
API Gateway	FastAPI
Monitoring	Prometheus + Grafana

3.3 Voice AI Module (Optional Add-On)

For clinics & service centers:

- SIP server (Asterisk)
- SIP trunk integration
- Whisper (STT)
- LLM inference
- TTS engine
- Call routing automation

This enables:

✓ AI receptionist
✓ Appointment scheduling
✓ FAQ handling
✓ Multilingual response

---

4. Deployment Phases

Phase 1 – Pre-Deployment Assessment

✓ Site network audit
✓ Bandwidth test (min 200 Mbps recommended)
✓ Power redundancy check
✓ Data sensitivity classification
✓ Regulatory compliance check (HIPAA if medical)

Deliverable:
Deployment Readiness Report

Phase 2 – Physical Installation

1. Rack or secure tower placement
2. UPS installation
3. Network connection
4. Static IP configuration
5. Secure VLAN setup (recommended)

Time estimate: 2–4 hours

Phase 3 – Software Installation

1. Install OS
2. GPU driver + CUDA
3. Docker installation
4. Deploy container stack
5. Configure monitoring
6. Test inference performance

Validation checklist:

✓ GPU recognized
✓ Test model inference

✓ Voice module test call
✓ Dashboard accessible

Phase 4 – AI Model Configuration

- Upload client-specific knowledge base
- Configure vector indexing
- Fine-tune prompt templates
- Set escalation rules
- Configure multilingual support

Phase 5 – Integration

Depending on client type:

Clinic

- Connect to scheduling software
- Integrate with CRM
- SIP trunk mapping

Travel Agency

- Booking system API integration
- Internal knowledge RAG
- Email automation

Law Firm

- Document indexing
- Case search AI
- Secure access control

---

5. Security & Compliance Standards

5.1 System Hardening

- SSH key-based authentication only
- Firewall rule restriction
- Fail2Ban
- Automatic security updates

5.2 Data Protection

- Local encrypted storage
- Encrypted backups
- VPN-only remote access
- Role-based access control

5.3 Compliance Considerations

Medical:

- HIPAA-compliant configuration
- Access logging enabled

Enterprise:

- SOC2-ready logging
- Audit trail retention

---

6. Maintenance Protocol

Daily

- GPU load monitoring
- Service health check

Weekly

- Backup verification
- Log review

Monthly

- Security patching
- Model performance review
- Update knowledge base

Quarterly

- Capacity planning
- Upgrade evaluation

---

7. Remote Management Architecture

PanAuro Central Management Console:

- Remote SSH tunnel
- Fleet monitoring dashboard
- Model update push system
- Performance analytics

This enables:

> Centralized control of distributed Edge Nodes.

---

8. Standard Performance Targets

Metric	Target
LLM Response	< 2 seconds
Voice Latency	< 1.5 seconds
Uptime	99.5%
GPU Utilization	40–70% optimal

---

9. Deployment SKU Strategy (For Channel Partners)

You should standardize 3 SKUs:

1. Edge Basic
2. Edge Pro
3. Edge Enterprise

This allows GTT Global and others to sell clearly defined packages.

---

10. Business Positioning Statement

You can present it as:

> "PanAuro operates a distributed Edge AI infrastructure layer connected to our international AI Factory, enabling enterprises to deploy private, secure, GPU-accelerated AI locally."

That sounds institutional.