CELLULE-PRO DOCUMENTATION

Technical Architecture

4-layer model · federation · MoE sharding

cellule.ai · pro.cellule.ai
2026-05-18

Contents

0.1 Cellule PRO — Reference architecture

This document is the source of truth for a Cellule PRO deployment at an enterprise customer.

The two-sentence pitch (by segment)

General pitch

For any company with dormant compute (laptops, desktops, idle servers):

> Companies own a huge amount of unused compute power. All it takes is to put it to work and bring in NPU servers to balance inference, allowing them to use LLMs securely and enjoy comfortable inference.

Targeted pitch — pro studios (VFX, design, cinema, scientific labs)

When the prospect already has high-end GPU workstations (RTX 4090/A6000, Radeon Pro W7900, Apple M3 Ultra, Quadro):

> You have the power, we have the solution!

The idea: no need to invest in a dedicated NPU server. The studio's GPU workstations, idle at night and on weekends, already provide 10-50× more capacity than a consumer NPU. Cellule PRO supports CUDA, ROCm, Vulkan, Metal and MLX to leverage the entire heterogeneous fleet.

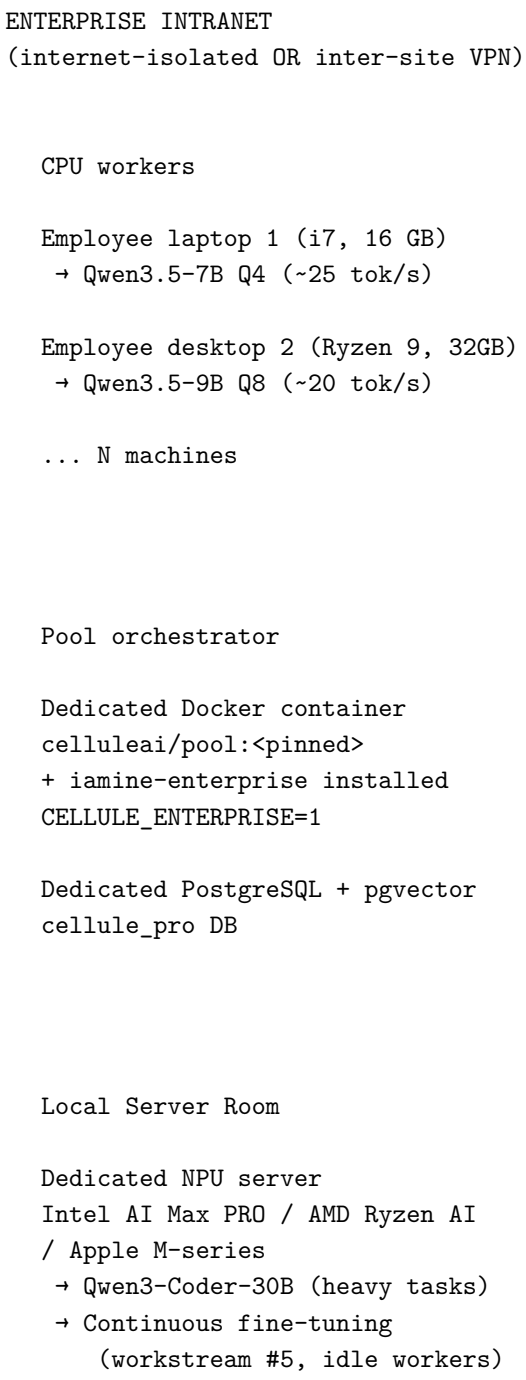
What this means concretely

1. **The existing resource pool** — in any company:
   - Employee laptops idle ~18h/day (nights, weekends, breaks)
   - CAD/dev desktops and workstations that idle in the evening
   - Underused backup or development servers
   - Graphics workstations at night This compute power is ALREADY paid for (hardware amortized, electricity, maintenance) but not put to use.
2. **The technology contribution** — a dedicated NPU server (Intel AI Max PRO, AMD Ryzen AI, Apple M-series, Jetson) installed in the customer's server room, which:
   - Serves heavy inference tasks (30B+ models) that cannot fit on a CPU laptop
   - Runs continuous fine-tuning (workstream #5) on the company's data
   - Balances the load between CPU employee workers and centralized compute
3. **The result** — an LLM inference infrastructure:
   - **Secure**: everything stays in the customer's intranet, nothing leaks to the internet
   - **Comfortable**: inference quality equivalent to ChatGPT/Claude thanks to the CPU + NPU mix
   - **Economical**: no massive GPU purchase, leverages existing hardware + a single dedicated NPU server
   - **Controlled**: fixed license, no per-token billing

Sales argument vs US competition

Instead of sending your data to a US provider at €20/employee/month (ChatGPT Enterprise, Claude for Business), wake up your existing machines and buy a single NPU server (€5-15k one-shot). Manage everything yourself. Zero data leak, predictable cost, native compliance.

Typical deployment topology



```
   Internal web UI     Multi-site VPN
(accessible from any    (additional
 authenticated           offices)
  employee endpoint)
```

Detailed components

1. CPU workers — employee machines

Hardware prerequisites: - Modern CPU: Intel i5/i7/i9 12th gen+, AMD Ryzen 5/7/9 series 5000+, Apple Silicon M1+ - RAM: 16 GB minimum, 32 GB recommended - Disk: 30 GB available (GGUF models) - OS: Linux / Windows / macOS

Default model: Qwen3.5-7B or 9B in Q4/Q8 depending on available RAM.

Employee install: - systemd service (Linux) / LaunchAgent (macOS) / scheduled task (Windows) - Starts at boot, registers with the intranet pool orchestrator - No manual config (auto-discovery via mDNS broadcast or IT-deployed config)

Resource policy (IT-configurable): - Worker runs at `nice=10` (low priority) - Pauses if CPU > 80% for 60s (avoid slowing down the employee's work) - Resumes when idle (CPU < 20%)

2. Dedicated NPU server

Hardware prerequisites: - Machine with integrated NPU (Intel AI Max PRO, AMD Ryzen AI, Apple M3 Ultra, Jetson) - 64 GB RAM minimum, 128 GB recommended - 500 GB NVMe SSD - Gigabit network (10 Gbit if possible)

Models served: - Qwen3-Coder-30B (code, complex reasoning) - Qwen3.5-35B-A3B (MoE, extended tasks) - Optional: model fine-tuned on enterprise data (workstream #5)

Role in the pool: - "Premium" worker auto-selected for tasks tagged `complex=True` - Fallback for CPU workers when p95 latency exceeds SLA - Host for continuous fine-tuning (night/weekend) via idle CPU workers contributing gradients

3. Pool orchestrator

Deployment: - Docker container `celluleai/pool:<pinned-version>` deployed on a dedicated Linux VM - `iamine-enterprise` package installed via `pip install` from a private wheel provided by CELLULE - `CELLULE_ENTERPRISE=1` set in the environment - PostgreSQL + pgvector in a container or adjacent VM

Config differences vs public cellule.ai: - `POOL_URL=http://pool.intranet.client.local:8080` (internal URL) - `FEDERATION_ENABLED=false` by default (mono-site) or `true` + intranet peers only - No outbound connection to public cellule.ai - `REGISTRATION_ENDPOINT=disabled` (no public sign-up)

Access: - Web UI via VPN or in the intranet only - Enterprise SSO (LDAP, Azure AD, Okta) — to be integrated in a later iteration - API tokens managed by the IT admin (not user self-serve)

4. Database

PostgreSQL 16 with `pgvector` and `pgcrypto` extensions.

Stored data: - Conversations encrypted with zero-knowledge Fernet (per user session) - Vectorized `user_memories` (RF=3 if multi-site via workstream #1) - `worker_reputation` + `inference_verifications` (workstream #4) - `agent_observations`, `agent_episodes` (public M13)

Backup: daily snapshots, 30-day retention, encrypted offsite (intranet NAS).

Sizing guide — how many workers for N employees?

Assumption: average usage = 20 requests/day/employee, target p50 latency < 3s.

Active employees	CPU workers	NPU server	Pool orchestrator RAM	Postgres
10-30	5-15 idle machines	1× mid-tier NPU (128 GB)	16 GB	50 GB SSD
30-100	15-50	1× high-end NPU or 2× mid	32 GB	200 GB SSD
100-300	50-150	2× high-end NPUs	64 GB	500 GB NVMe SSD
300+	federated multi-site	2+ NPUs per site	128 GB	1 TB NVMe + read replicas

A modern CPU worker serves ~5-10 requests/minute in support (20-40 tok/s with a 7-9B model). So 1 worker = ~300-600 requests/hour.

Customer onboarding process (9 steps)

1. **Kickoff** — 1h call to establish: employee count, compliance constraints, sites, target models
2. **Infra audit** — the customer's IT lists machines + OS + network topology
3. **Quote + contract** — license + setup + support + NDA
4. **Hardware order if needed** — NPU server if the customer doesn't have one
5. **Install pool orchestrator** — docker container on customer-provided Linux VM
6. **Worker deployment** — IT script to install the agent on endpoints (MDM, GPO, bash)
7. **IT admin training** — 2h: monitoring, upgrade, backup, troubleshooting
8. **Employee training** — 30 min: how to access the web UI, where their conversations live
9. **Acceptance + production rollout** — SLA tests over 2 weeks before switching

Total duration: 2-6 weeks depending on size.

Native compliance

- **GDPR art. 32**: at-rest encryption (Fernet), in-transit encryption (intranet TLS), pseudonymization (token_hash)
- **HIPAA §164.308(a)(7)**: contingency plan = backup + documented RTO
- **SOC 2 Type II**: audit trail (workstream #4 proof of inference + reputation + max-determinism mode)
- **ISO 27001**: documented ISMS, access controls, backup, incident response

Structural differences vs public cellule.ai

Aspect	Community Cellule	Cellule PRO
Domain	cellule.ai	customer intranet (e.g. ai..local)
Workers	Z2, Gladiator, master.86 (David's infra)	Customer employee machines
Users	Anon + free account	Authenticated enterprise employees
Data location	Contabo Nuremberg VPS	Customer intranet
Code	`CELLULEAI/POOL` AGPLv3	`CELLULEAI/POOL` + private `iamine-enterprise`
Enterprise plugins	Never loaded	All loaded (CELLULE_ENTERPRISE=1)
Fine-tuning on customer data	No	Yes (workstream #5, idle workers)
Pricing	Free	Fixed license + setup + support
SLA	Best-effort	Contractual

Takeaway

Cellule PRO is a distinct product offer, not a premium version of public cellule.ai. It shares the open-source technical foundation (AGPLv3) and adds a suite of proprietary plugins under commercial license + NDA.

The business model is **license + service**, not pay-per-token. The customer owns their infrastructure, their data, their control. CELLULE provides the software and the support.