Auri
Local Enterprise Intelligence

TECHNISCHE SERVERBESCHREIBUNG

ON-PREMISE INFRASTRUKTUR

Ihre Hardware.
Ihre Kontrolle.
Ihr Auri-Server.

Dieses Dokument beschreibt die technische Server-Konfiguration für den lokalen Betrieb von Auri – von der Hardware-Auslegung über den Software-Stack bis zu Netzwerk und Betrieb. Alles läuft in Ihrem Rechenzentrum. Alles bleibt unter Ihrer Kontrolle.

Version	Umfang	Sprache
2026	Technische Beschreibung zur Auri-Webseite	Deutsch

01 · ARCHITEKTUR

Die Auri-Server-Architektur

Auri läuft als dedizierter GPU-Server in Ihrem eigenen Rechenzentrum. Das Sprachmodell und alle Daten verbleiben lokal – es ist keine Internetverbindung erforderlich. Die Architektur ist bewusst einfach gehalten: ein Server, der das Modell hostet und Ihre Systeme kontrolliert anbindet.

Je nach Teamgröße und Anforderungen wird die Konfiguration von einem kompakten Ein-Nutzer-System bis zu einem mehrstufigen Business-Setup mit mehreren GPUs skaliert. Die Grundprinzipien bleiben dabei identisch.



GPU-Server vor Ort

Ein dedizierter Server in Ihrem Rechenzentrum hostet das Modell und die Anwendung – physisch bei Ihnen.



LLM lokal

Die Inferenz läuft über llama.cpp oder vLLM direkt auf Ihrer Hardware – ohne Cloud-Runden.



Kein Internet nötig

Auri ist auch in einem air-gapped Netzwerk vollständig betreibbar. Keine externe Abhängigkeit.

Der entscheidende Punkt: Auri ist **keine Cloud-Lösung**. Modell, Daten und Betrieb liegen vollständig in Ihrem Verantwortungsbereich. Sie entscheiden über Hardware, Netzwerk und Zugriff – Auri fügt sich in Ihre bestehende IT ein.

02 · HARDWARE · PRIVAT

Konfiguration – Privat

Die Privat-Konfiguration ist für den Betrieb mit einem Nutzer ausgelegt. Sie bietet schnelle Antwortzeiten und eine stabile Basis für den 24/7-Betrieb – kompakt, effizient und wirtschaftlich.

KOMPONENTE	SPEZIFIKATION	RICHTPREIS
GPU	RTX 4090 / 5090 (ab 24 GB VRAM)	ca. €1.500 – 2.500
CPU	Ryzen 9 / Core i9 (ab 8 Kerne)	ca. €400 – 600
Arbeitsspeicher	64 GB DDR5 (Standard, bis 128 GB)	ca. €200 – 350
Speicher	2 TB NVMe Gen4 (Standard, bis 4 TB)	ca. €200 – 400
Netzwerk	2.5 GbE (Standard)	ca. €100 – 200

Hinweis: Die genannten Preise beziehen sich ausschließlich auf die Hardware-Komponenten (Richtwerte, Stand 2026). Die Kosten für Auri selbst – Lizenz, Setup, Integration und Support – ermitteln wir individuell in einem persönlichen Termin.

03 · HARDWARE · BUSINESS

Konfiguration – Business

Die Business-Konfiguration ist für mehrere parallele Nutzer und höhere Durchsatzraten ausgelegt. Mehrere GPUs, ECC-Arbeitsspeicher und RAID-Speicher sorgen für Skalierbarkeit, Stabilität und Redundanz.

KOMPONENTE	SPEZIFIKATION	RICHTPREIS
GPU	1–4× RTX 5090 (bis 512 GB VRAM gesamt)	je nach Auslegung
CPU	Threadripper / Xeon (ab 16 Kerne, bis 64+)	ca. €800 – 2.500
Arbeitsspeicher	128–512 GB DDR5 ECC (bis 1 TB)	ca. €500 – 2.000
Speicher	RAID, 4–8 TB NVMe Gen4 (bis 32 TB)	ca. €500 – 2.000
Netzwerk	10 GbE (Standard ab Mittel-Tier)	ca. €200 – 500

Die Business-Konfiguration ist modular aufgebaut: Sie beginnt mit einer einzelnen RTX 5090 und wächst schrittweise – zusätzliche GPUs, mehr Speicher oder ein RAID-Array, sobald der Bedarf steigt. So investieren Sie nur, was Ihr Team tatsächlich braucht.

04 · KOMPONENTEN

Die Komponenten im Detail

Jede Komponente des Auri-Servers hat eine klare Aufgabe. Zusammen ergeben sie ein System, das das Sprachmodell schnell lädt, stabil betreibt und Ihre Systeme zuverlässig anbindet.

GPU – das Herzstück

Führt die Inferenz des Sprachmodells aus und bestimmt Geschwindigkeit und Qualität der Antworten. Optimiert für Qwen-Modelle; bei Business-Setups parallel über mehrere GPUs.

CPU – Orchestrierung

Steuert das System, verarbeitet Anfragen und koordiniert die Datenflüsse zwischen GPU, Speicher und Schnittstellen. Stabile Basis für den 24/7-Betrieb.

Arbeitsspeicher

Hält aktive Daten und Modell-Teile für schnelle Zugriffe bereit. Bei Business-Setups mit ECC für Fehlerkorrektur bei hohem parallelen Bedarf.

Speicher (NVMe)

Beherbergt das KI-Modell, Ihre Daten und die Anwendung. NVMe Gen4 sorgt für schnelle Ladezeiten; RAID bietet Redundanz.

Netzwerk

Verbindet Auri mit Ihren bestehenden Systemen – ERP, Microsoft 365, Datenbanken und Dateien. Sichere lokale Kommunikation, keine externe Datenablage.

Gehäuse & Betrieb

Server-Gehäuse für den Rack- oder Standalone-Betrieb im Rechenzentrum. Ausgelegt für kontinuierlichen Betrieb und einfache Wartung.

05 · SOFTWARE & BETRIEB

Software-Stack & Betrieb

Auf der Hardware läuft ein schlanker, lokal betriebener Software-Stack. Die Inferenz des Sprachmodells erfolgt direkt auf der GPU – ohne externe Dienste. Der Betrieb ist auf Stabilität und Nachvollziehbarkeit ausgelegt.

- ✓ **Inferenz-Engine** – llama.cpp oder vLLM führen das Sprachmodell lokal auf der GPU aus. Beide sind etablierte, offene Laufzeitumgebungen für große Sprachmodelle.
- ✓ **Sprachmodell** – Auri ist auf Qwen-Modelle optimiert. Das Modell wird lokal geladen und bleibt auf Ihrem Server.
- ✓ **24/7-Betrieb** – Der Server ist für den kontinuierlichen Betrieb ausgelegt. Modell-Loading, Anfragen und Datenflüsse laufen stabil im Hintergrund.
- ✓ **Kontrollierte Anbindung** – Ihre Systeme (ERP, Microsoft 365, Datenbanken, Dateien) werden kontrolliert und nachvollziehbar angebunden – nur mit den Daten, die Sie freigeben.
- ✓ **Keine externe Abhängigkeit** – Der Betrieb funktioniert auch ohne Internetverbindung. Updates und Wartung erfolgen nach Ihrem Zeitplan.

06 · NETZWERK & SICHERHEIT

Netzwerk & Sicherheit

Auri ist so konzipiert, dass Daten und Betrieb innerhalb Ihres Netzwerks bleiben. Die Anbindung an Ihre Systeme erfolgt lokal und kontrolliert – ohne dass Informationen Ihr Netzwerk verlassen.

Lokale Kommunikation

Auri kommuniziert ausschließlich mit Systemen in Ihrem Netzwerk. Es findet keine externe Datenablage statt.

Kontrollierter Zugriff

Sie entscheiden, welche Systeme und Daten Auri nutzen darf. Der Zugriff ist kontrolliert und nachvollziehbar.

Air-gapped Betrieb

Auf Wunsch läuft Auri vollständig ohne Internetverbindung – ideal für besonders schutzbedürftige Umgebungen.

Datenhoheit

Modell und Daten liegen physisch bei Ihnen. Volle Souveränität, volle Kontrolle – jederzeit.

Die konkrete Netzwerkkonfiguration – Firewall-Regeln, Segmentierung, Zugriffsberechtigungen – stimmen wir im Projekt auf Ihre bestehende IT ab. So fügt sich Auri sicher in Ihre vorhandene Sicherheitsarchitektur ein.

07 · ZUSAMMENFASSUNG

Zusammenfassung & Nächste Schritte

Die Konfiguration im Überblick

	PRIVAT	BUSINESS
GPU	1× RTX 4090/5090	1–4× RTX 5090
CPU	Ryzen 9 / i9 (8+ Kerne)	Threadripper / Xeon (16+ Kerne)
RAM	64 GB DDR5	128–512 GB DDR5 ECC
Speicher	2 TB NVMe Gen4	4–8 TB NVMe (RAID)
Netzwerk	2.5 GbE	10 GbE
Nutzer	1	mehre parallel

Nächste Schritte

✓ **Bedarfsanalyse** – Wir ermitteln Teamgröße, Systeme und Anforderungen für die passende Konfiguration.

✓ **Hardware-Auslegung** – Sie erhalten eine konkrete Server-Konfiguration mit Komponenten und Richtwerten.

✓ **Aufbau & Pilot** – Auri wird aufgestellt und an einem konkreten Anwendungsfall erprobt.