

Leistungsbeschreibung

für IT-Infrastruktur zum lokalen Betrieb großer Sprachmodelle (drei Lose)

Inhalt

1	Präambel	2
1.1	Begriffsklärung	2
1.2	Vertragsart	2
2	Einleitung	2
3	Allgemeine Bestimmungen	3
3.1	Anforderungen an den Auftragnehmer	3
3.2	Angebotsabgabe	3
3.4	Lieferung und Liefersicherheit	3
4	Los 1: GPU-Server für KI-Inferenz.....	4
4.1	Zweck und Systemübersicht	4
4.2	Konfiguration – technische Anforderungen.....	4
4.3	Erweiterbarkeit	5
4.4	Installation und Inbetriebnahme	5
4.5	Optionen	5
5	Los 2: Netzwerkspeichersystem (NAS).....	6
5.1	Zweck und Systemübersicht	6
5.2	Technische Anforderungen	6
5.3	Optionen	6
6	Los 3: Web-Server (öffentliche Webdienste).....	6
6.1	Zweck und Systemübersicht	6
6.2	Technische Anforderungen	6
6.3	Optionen	7
7	Sonstiges	7
7.1	Wartung, Support und Gewährleistung	7
7.2	Anzugebende Nachweise (je Los)	7
7.3	Preisobergrenze	8
8	Vorhandene Technik und Infrastruktur	8

1 Präambel

1.1 Begriffsklärung

Zur Klärung der in der nachfolgenden Leistungsbeschreibung genutzten Begriffe sowie die generell gültigen Regeln werden diese hier kurz erläutert:

- **Auftraggeber:** Auftraggeber dieser Leistungsbeschreibung ist die Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU).
- **Auftragnehmer bzw. Bieter:** Auftragnehmer oder Bieter sind gleichwertige Begriffe und beschreiben das Unternehmen, das an der öffentlichen Ausschreibung teilnimmt. Die angebotenen Produkte müssen nicht vom Auftragnehmer sein, sondern können von anderen Herstellern sein.
- **Hersteller:** Der Hersteller ist das Unternehmen, das die Produkte des Angebotes herstellt. Der Auftragnehmer muss mit den Produkten vertraut sein oder einen Kontakt zum Hersteller haben, um Rückfragen beantworten zu können.
- **Dritte:** Als Dritter bezeichnet diese Ausschreibung Unternehmen, die vom Auftragnehmer unterbeauftragt werden. Dritte müssen durch den Auftragnehmer vorab mit Namen und Ansprechpartner benannt werden.
- **Los:** Die Beschaffung ist in drei eigenständige Lose gegliedert. Angebote können für ein Los, für mehrere oder für alle Lose abgegeben werden. Wertung und Zuschlag erfolgen je Los getrennt; ein Anspruch auf den Zuschlag mehrerer Lose an denselben Bieter besteht nicht.

Die Projektsprache, in der die Kommunikation und Dokumentation im Projekt durchgeführt wird, ist generell Deutsch, aber eingeschränkt auch in Englisch möglich.

1.2 Vertragsart

Es wird ein Vertragsabschluss auf Basis des EVB-IT Kaufvertrags in Langfassung inklusive der dazugehörenden AGBs angestrebt. Allgemeine Geschäftsbedingungen des Bieters werden nicht Vertragsbestandteil; entgegenstehende oder abweichende Bedingungen des Bieters (insbesondere freibleibende Preise oder unverbindliche Liefertermine) sind unbeachtlich.

2 Einleitung

Die vorliegende Leistungsbeschreibung spezifiziert die IT-Infrastruktur für den lokalen Betrieb großer Sprachmodelle (KI-Inferenz), die im Rahmen des Projekts Kompetenzzentrum für smarte Konnektivität (im Folgenden KSK) am Lehrstuhl Funkkommunikation und Navigation der RPTU Kaiserslautern-Landau beschafft wird. Das KSK ist ein vom Bundesministerium für Digitales und Staatsmodernisierung gefördertes Forschungsprojekt und fungiert als hersteller- und technologie neutrale Test- und Transferplattform. Ziel ist es, die Einführung wirtschaftlich tragfähiger digitaler Konnektivitätslösungen in Deutschland zu beschleunigen und kleine und mittlere Unternehmen bei Auswahl, Bewertung und Skalierung von Konnektivitätstechnologien neutral und anwenderorientiert zu unterstützen. Die Projektlaufzeit endet am 31. Dezember 2029.

Kern der Beschaffung ist ein GPU-Server als Inferenz-Backend eines Retrieval-Augmented-Generation-Systems (RAG), das Anwenderinnen und Anwender bei der Auswahl geeigneter Funktechnologien unterstützt und Antworten ausschließlich auf Grundlage geprüfter, zitierfähiger Quellen erzeugt. Der Betrieb erfolgt vollständig lokal auf eigener Hardware; eine Übertragung von Daten an externe Dienste

findet nicht statt. Ergänzend werden ein Netzwerkspeichersystem für Datenablage und Sicherungen sowie ein Web-Server für die öffentlich erreichbaren Dienste beschafft. Die Beschaffung erfolgt in drei Losen:

- **Los 1:** GPU-Server für KI-Inferenz (teilbestückt, erweiterbar auf den Vollausbau)
- **Los 2:** Netzwerkspeichersystem (NAS)
- **Los 3:** Web-Server (1 HE) für die öffentlich erreichbaren Webdienste

3 Allgemeine Bestimmungen

3.1 Anforderungen an den Auftragnehmer

Der Auftragnehmer muss mit den angebotenen Produkten vertraut sein oder über einen direkten Kontakt zum Hersteller verfügen, um technische Rückfragen beantworten und Konfigurations- sowie Anwendungsfragen während der Projektlaufzeit unterstützen zu können.

3.2 Angebotsabgabe

Für die Abgabe von Geboten gelten folgende Rahmenbedingungen:

- Bei allen Komponenten muss es sich um Fabrikneuwere handeln. Generalüberholte, wiederaufbereitete (refurbished), reparierte oder gebrauchte Geräte sowie Vorführ-, Ausstellungs- oder Demogeräte sind unzulässig und führen zum Ausschluss des Angebots. Die Ware muss aus dem offiziellen Vertriebskanal des Herstellers für den Europäischen Wirtschaftsraum stammen und die uneingeschränkte Herstellergarantie umfassen; auf Verlangen ist die Bezugsquelle nachzuweisen.
- Die angebotenen Preise sind Festpreise in Euro (netto, zuzüglich gesetzlicher Umsatzsteuer) und gelten unverändert bis zur vollständigen Vertragserfüllung einschließlich der vereinbarten Lieferfrist. Preisvorbehalte, Preisgleit- oder Preisanpassungsklauseln sowie Währungs- oder Zollvorbehalte sind unzulässig. Die Bieter sind an ihr Angebot bis zum Ablauf von 90 Kalendertagen nach Ablauf der Angebotsfrist gebunden.
- Soweit in dieser Leistungsbeschreibung Produkte, Verfahren oder Kennwerte genannt werden, gilt stets „oder gleichwertig“; die Gleichwertigkeit ist vom Bieter nachzuweisen.
- Die Bezahlung erfolgt nach endgültiger gemeinsamer Abnahme des jeweiligen Loses durch Auftraggeber und Auftragnehmer.
- Je Los sind der Gesamtpreis der Hauptkomponenten anzugeben. Die in den Losbeschreibungen genannten Optionen sind gesondert auszureisen.

3.4 Lieferung und Liefersicherheit

Die Lieferung der im Folgenden beschriebenen Komponenten sollte schnellstmöglich erfolgen, spätestens jedoch 120 Kalendertage nach Zuschlagserteilung. Der Bieter muss bei Abgabe des Angebotes eine zuverlässige Angabe zu den Lieferzeiten der beschriebenen Hardware machen; absehbare Lieferrisiken sind offenzulegen. Bei voraussichtlicher Nichteinhaltung der vereinbarten Liefertermine ist der Auftraggeber unverzüglich zu benachrichtigen und eine unbegrenzte Bankbürgschaft in Höhe der nicht gelieferten Komponenten zu stellen. Die Laufzeit zeitlich begrenzter Lizenzen sowie etwaiger Wartungs-, Pflege- und Supportzeiträume beginnt erst mit der erfolgreichen

Abnahme bzw. Inbetriebnahme des Gesamtsystems, nicht bereits mit Lieferung oder Aktivierung einzelner Komponenten.

Zahlung im Haushaltsjahr 2026 gegen Sicherheitsleistung: Kann die Lieferung eines Loses nicht bis zum 31.12.2026 erfolgen, ist der Auftraggeber berechtigt, den Auftragswert des betroffenen Loses ganz bereits im Jahr 2026 zu zahlen, gegen Übergabe einer unbefristeten Bankbürgschaft. Die Bürgschaft wird nach vollständiger, vertragsgerechter Lieferung und Abnahme zurückgegeben. Die Inanspruchnahme dieser Regelung steht im Ermessen des Auftraggebers; die Bezahlung nach Abnahme (Abschnitt 3.2) bleibt im Übrigen unberührt.

Lieferungen müssen dem Auftraggeber eine Woche vorher angekündigt werden. Die Lieferung erfolgt frei Haus (verzollt und versichert) einschließlich Transport bis zum Aufstellungsort (Serverraum); sämtliche Transport-, Versand- und Versicherungskosten trägt der Auftragnehmer. Einzelne Systeme können mehr als 60 kg wiegen; der Auftragnehmer stellt geeignete Transportmittel und ausreichend Personal bereit. Die Gefahr des zufälligen Untergangs und der zufälligen Verschlechterung geht erst mit der Anlieferung bzw. der gemeinsamen Abnahme auf den Auftraggeber über, Transportschäden gehen zu Lasten des Auftragnehmers. Die Adresse für die Anlieferung ist:

RPTU Kaiserslautern-Landau

Lehrstuhl Funkkommunikation und Navigation

Paul-Ehrlich-Straße 11,

67663 Kaiserslautern

4 Los 1: GPU-Server für KI-Inferenz

4.1 Zweck und Systemübersicht

Gegenstand von Los 1 ist ein dedizierter GPU-Server als Inferenz-Backend des in Abschnitt 2 beschriebenen RAG-Systems. Der Server wird teilbestückt beschafft und muss ohne Austausch von Kernkomponenten (Gehäuse, Netzteile, Kühlung, Riser, Verkabelung) auf den Vollausbau mit acht Grafikprozessoren aufgerüstet werden können.

4.2 Konfiguration – technische Anforderungen

- **Bauform:** 19-Zoll-Rackserver, maximal 5 Höheneinheiten, Luftkühlung, einschließlich Einbauschienen
- **Prozessoren:** Zwei Server-Prozessoren (x86-64) mit zusammen mindestens 128 physischen Kernen. Die Rechenleistung ist über den Benchmark SPECrate2017_int_base nachzuweisen; gefordert ist ein Wert von mindestens 1.500 für eine Zwei-Sockel-Konfiguration mit dem angebotenen Prozessormodell.
- **Arbeitsspeicher:** 768 GB DDR5-RDIMM, bestückt als 24x 32 GB mit 6400 MT/s
- **Grafikprozessoren (Erstbestückung):** 4x NVIDIA RTX PRO 6000 Blackwell Server Edition, je 96 GB GDDR7 ECC, Betrieb mit 600 W je Karte, einschließlich der werkseitigen GPU-Verkabelung für den 600-W-Betrieb. Die Festlegung auf dieses Modell ist durch den Auftragsgegenstand bedingt, da der eingesetzte Software-Stack (vLLM auf Basis von NVIDIA CUDA) auf die Blackwell-Architektur mit hardwareseitiger Unterstützung der Rechenformate FP8 und FP4 abgestimmt ist

und für die vorgesehene Erweiterung auf acht Grafikprozessoren baugleiche Karten erforderlich sind.

- **GPU-Kapazität des Gehäuses:** insgesamt mindestens 8 Dual-Width-PCIe-Gen5-x16-Steckplätze, jeder für 600-W-Grafikprozessoren ausgelegt und mit elektrischer PCIe-5.0-x16-Anbindung; alle acht Steckplätze sind mit Risern und GPU-Stromversorgungskabeln für den 600-W-Betrieb vorbestückt
- **Datenspeicher:** 2x 7,68 TB NVMe-SSD (E3.S oder U.2, PCIe Gen5, leseoptimiert, Enterprise-Klasse); Gehäuse mit mindestens 8 NVMe-Einschüben
- **Boot-Medium:** separates, gespiegeltes Boot-Medium mit 2x 960 GB M.2-NVMe im Hardware-RAID 1, getrennt vom Datenspeicher
- **Netzwerk:** mindestens 2x 25 GbE; eine Ausführung mit 2x 100 GbE (QSFP) ist als Option gesondert auszureisen
- **Stromversorgung:** redundante Hot-Plug-Netzteile der Effizienzklasse Titanium, Eingang 200–240 V, bereits für den Vollausbau mit 8 Grafikprozessoren dimensioniert (Referenz: 8x 3.200 W in 4+4-Redundanz); Netzanschlusskabel C19/C20 sind mitzuliefern
- **Management:** dedizierte Out-of-Band-Fernverwaltung (BMC) mit IPMI und/oder Redfish einschließlich der erforderlichen Lizenz
- **Betriebssystem:** ohne Betriebssystem; nachgewiesene Herstellerfreigabe für Ubuntu Server LTS

4.3 Erweiterbarkeit

- Die Erweiterung von vier auf acht Grafikprozessoren muss ohne Austausch von Netzteilen, Kühlung, Risern oder Verkabelung möglich sein; Stromversorgung und Kühlung sind bereits bei Lieferung für den Vollausbau auszulegen.
- Eine spätere Erweiterung des Arbeitsspeichers muss möglich sein; der Bieter beschreibt die empfohlene Modulbestückung ohne Verlust der Speicherbandbreite.
- Mindestens sechs NVMe-Einschübe müssen für spätere Erweiterungen frei bleiben.

Klarstellung: Etwaige spätere Erweiterungen (zusätzliche Grafikprozessoren, Arbeitsspeicher, Datenträger) sind nicht Gegenstand dieses Auftrags und würden bei Bedarf gesondert beschafft. Sofern der Bieter Richtpreise für Erweiterungen angibt, erfolgt dies rein informatorisch und unverbindlich; ein Optionsrecht des Auftraggebers, eine Bezugsverpflichtung oder ein Anspruch des Auftragnehmers werden hierdurch nicht begründet.

4.4 Installation und Inbetriebnahme

Der Auftragnehmer liefert den Server bis in den Serverraum des Auftraggebers und übergibt mit der Lieferung sämtliche Zugangsdaten sowie die zur Rackmontage erforderlichen Einbauschienen und Anschlusskabel. Einbau, Anschluss an die Stromversorgung und Inbetriebnahme einschließlich der Installation von Betriebssystem und Anwendungssoftware erfolgen durch den Auftraggeber.

4.5 Optionen

Optionen werden nur zusammen mit dem Grundgerät vergeben und sind gesondert auszureisen.

- **Option 1 – Netzwerk 100 GbE:** Ausführung bzw. Nachrüstsatz mit 2x 100 GbE (QSFP); Aufpreis ist anzugeben.

- **Option 2 – Supportverlängerung:** Verlängerung des Hardware-Supports gemäß Abschnitt 7.1 auf 60 Monate; Aufpreis ist anzugeben.

5 Los 2: Netzwerkspeichersystem (NAS)

5.1 Zweck und Systemübersicht

Gegenstand von Los 2 ist ein Rack-Netzwerkspeichersystem für die interne Datenablage (Rohdaten der Wissensbasis, Modell-Artefakte) und als Ziel für Datensicherungen der Server, sowie für den internen Betrieb einer Nextcloud-Instanz zur Dateiablage und zum Dateiaustausch innerhalb des Hochschulnetzes einschließlich der zugehörigen Sicherungen. Das System wird ausschließlich im internen Netz betrieben.

5.2 Technische Anforderungen

- **Bauform:** 19-Zoll-Rackmontage, maximal 3 Höheneinheiten, einschließlich Einbauschiene
- **Einschübe:** mindestens 8 Hot-Swap-Einschübe (3,5"/2,5")
- **Erstbestückung:** 4 Datenträger der Enterprise-Klasse (24/7-Betrieb, für NAS-/RAID-Einsatz freigegeben) mit je mindestens 12 TB; nutzbare Nettokapazität mindestens 20 TB bei Konfiguration mit doppelter Parität (RAID 6, RAID-Z2 oder gleichwertig); mindestens 4 Einschübe müssen für spätere Erweiterungen frei bleiben
- **Arbeitsspeicher:** mindestens 32 GB ECC-RAM
- **Netzwerk:** mindestens 2x 10 GbE (SFP+ oder 10GBASE-T)
- **Stromversorgung:** zwei Hot-Plug-Netzteile in 1+1-Redundanz, wobei jedes Netzteil die Last des Systems bei maximaler Bestückung allein tragen kann; Effizienzklasse mindestens 80 PLUS Gold; Eingangsspannung 230 V AC;

5.3 Optionen

- **Option 1 – Erweiterung:** Richtpreis je zusätzlichem Datenträger der angebotenen Baureihe (unverbindlich, rein informativ).
- **Option 2 – Supportverlängerung:** Verlängerung des Supports gemäß Abschnitt 7.1 auf 60 Monate; Aufpreis ist anzugeben.

6 Los 3: Web-Server (öffentliche Webdienste)

6.1 Zweck und Systemübersicht

Gegenstand von Los 3 ist ein kompakter Rack-Server für die öffentlich erreichbaren Webdienste des Projekts (Projekt-Website, interaktive Webplattform, Demonstrator-Zugänge). Auf dem System wird ein Webserver mit Reverse-Proxy (Apache) betrieben, der Anfragen aus dem Internet entgegennimmt und an die internen Dienste weiterleitet. Der Server ist das einzige von außen erreichbare System und wird in einem eigenen Netzsegment (DMZ) betrieben; er hält selbst keine schützenswerten Nutzdaten.

6.2 Technische Anforderungen

- **Bauform:** 19-Zoll-Rackserver, 1 Höheneinheit, einschließlich Einbauschiene

- **Prozessor:** ein Server-Prozessor (x86-64) mit mindestens 8 physischen Kernen und Unterstützung für DDR5-ECC-Speicher. Die Rechenleistung ist über den Benchmark SPECrate2017_int_base nachzuweisen; gefordert ist ein Wert von mindestens 80 für eine Ein-Sockel-Konfiguration mit dem angebotenen Prozessormodell. Der Nachweis erfolgt mit dem Angebot durch Angabe eines auf www.spec.org veröffentlichten Ergebnisses;
- **Arbeitsspeicher:** mindestens 64 GB ECC, erweiterbar auf mindestens 128 GB
- **Datenspeicher:** 2x mindestens 960 GB SSD (Enterprise-Klasse) im RAID 1 (hardware- oder softwareseitig)
- **Netzwerk:** mindestens 2x 1 GbE, getrennt nutzbar für externes und internes Netzsegment
- **Stromversorgung:** zwei Hot-Plug-Netzteile in 1+1-Redundanz, wobei jedes Netzteil die Last des Systems bei maximaler Bestückung allein tragen kann; Effizienzklasse mindestens 80 PLUS Platinum; Eingangsspannung 230 V AC; Geräteanschluss IEC C14, passende Netzkabel C13/C14 sind mitzuliefern;
- **Management:** dedizierte Out-of-Band-Fernverwaltung (BMC) mit IPMI und/oder Redfish einschließlich der erforderlichen Lizenz
- **Betriebssystem:** ohne Betriebssystem; nachgewiesene Herstellerfreigabe für Ubuntu Server LTS. Die Installation von Betriebssystem und Anwendungen (Apache, Webanwendungen) erfolgt durch den Auftraggeber.

6.3 Optionen

- **Option 1 – Supportverlängerung:** Verlängerung des Supports gemäß Abschnitt 7.1 auf 60 Monate; Aufpreis ist anzugeben.

7 Sonstiges

7.1 Wartung, Support und Gewährleistung

Die Gewährleistung muss mindestens 24 Monate auf das jeweilige Gesamtsystem betragen und ist im Angebot deutlich aufzuführen. Darüber hinaus ist je Los ein herstellergestützter Hardware-Support mit einer Laufzeit von mindestens 36 Monaten ab Abnahme anzubieten, mit Vor-Ort-Service und Reaktion am nächsten Arbeitstag einschließlich Ersatzteilversorgung.

Der Auftragnehmer muss über die gesamte Projektlaufzeit (bis mindestens 31.12.2029) technischen Support für die gelieferte Hardware bereitstellen. Der Support muss mindestens zu üblichen Geschäftszeiten über einen benannten Ansprechpartner per E-Mail und Telefon erreichbar sein; die Kommunikation sollte in deutscher, mindestens jedoch in englischer Sprache möglich sein. Etwaige Kosten für den Support während der Projektlaufzeit sind im Angebot anzugeben und in den Gesamtpreis einzurechnen.

7.2 Anzugebende Nachweise (je Los)

- Vollständige technische Spezifikation der angebotenen Produkte (Hersteller, Modellbezeichnung, Komponenten) sowie eine nachvollziehbare Bestätigung jeder Anforderung des jeweiligen Loses (Erfüllungsübersicht mit Verweis auf Datenblatt oder Herstellererklärung)
- Zuverlässige Angabe der Lieferzeiten sowie verbindlicher Liefertermin (Kalenderdatum) je Los

- Los 1: Leistungsaufnahme und Stromanschluss bei Vollausbau, Wärmeabgabe, Abmessungen, Gewicht und Einbautiefe; Hinweis auf eine gegebenenfalls erforderliche Endnutzererklärung (Exportkontrolle) für die Grafikprozessoren einschließlich der zugehörigen Formulare
- Los 2: Kapazität brutto und netto, vorgesehene RAID-Konfiguration, Datenträgertyp
- Los 3: Abmessungen, Einbautiefe und Leistungsaufnahme

7.3 Preisobergrenze

Für die einzelnen Lose gelten folgende Preisobergrenzen (brutto):

Los 1: GPU-Server für KI-Inferenz	275.000,00 €
Los 2: Netzwerkspeichersystem	11.000,00 €
Los 3: Web-Server	19.000,00 €

8 Vorhandene Technik und Infrastruktur

Am Aufstellungsort steht ein klimatisierter Serverraum mit einem 19-Zoll-Serverrack (Einbautiefe mindestens 900 mm, Gesamttiefe 1200mm) zur Verfügung. Die Stromversorgung des Racks (unterbrechungsfreie Stromversorgung und Stromverteilungen mit Abgängen IEC C19 und C13) wird bauseits bereitgestellt bzw. gesondert beschafft und ist nicht Gegenstand dieser Ausschreibung. Die Netzsegmentierung (DMZ, Firewall) erfolgt durch den Auftraggeber. Es handelt sich um eine Erstbeschaffung; vorhandene Systeme stellen keine Voraussetzung für den Betrieb der angebotenen Geräte dar.