



Gz.: VII 3 – 4.1.7. / 20-26

ANLAGE 2

Leistungsbeschreibung

„Leistungsbeschreibung für die Erweiterung der

Machine Learning Cloud Tübingen

am HLRS“

Auftraggeber

Eberhard Karls Universität Tübingen
Zentrale Verwaltung VII – Finanzen
Abteilung Einkauf
Geschwister-Scholl-Platz
D-72074 Tübingen

**EU-Ausschreibung - offenes Verfahren
nach § 119 GWB und VgV**



Inhaltsverzeichnis

1. EINLEITUNG	4
2. RAHMENBEDINGUNGEN	4
2.1. Infrastruktur und Betrieb	5
2.2. Begriffe, Abkürzungen und Definitionen	5
3. BEWERTUNGSKRITERIEN	6
3.1. Anforderungen der Kategorie A, Ausschlusskriterien (K.o.-Kriterien)	6
3.2. Anforderungen der Kategorie B, Bewertungskriterien	6
4. ORTSBESICHTIGUNG (VERPFLICHTEND):	6
5. TECHNISCHE SPEZIFIKATIONEN UND KONFIGURATION	7
5.1. Übersicht Bestandsystem	7
5.2. Allgemeine Anforderungen	8
5.3. Netzwerke	15
5.3.1. Management-Netzwerk	15
5.3.2. Provisionierungsnetz	16
5.3.3. InfiniBand Netzwerk	18
5.4. GPU - Rechenknoten	20
5.5. CPU - Rechenknoten	23
5.6. Kühlung	25
5.7. Server Rack	27
5.8. Facility Access & Service	29
5.9. Clustererweiterung	29



6. LIEFERUNG UND AUFBAU	29
6.1. Leistungsumfang, Abnahme und Endkontrolle	31
7. ANHANG	33
7.1. Technische Anschlussbedingungen für die Erweiterung des „Conway“ Clusters am HLRS	33



Leistungsbeschreibung:

1. Einleitung

Die Eberhard Karls Universität Tübingen beabsichtigt die Beschaffung eines High Performance Computing Clusters als Teil der ML Cloud Tübingen. Der anzuschaffende Cluster soll die am HLRS betriebene Installation „Conway“ um weitere Compute Ressourcen erweitern.

Das System soll hoch performante und verteilte KI Experimente für Wissenschaftler des Tübingen AI Centers, des Exzellenzclusters für Maschinelles Lernen in den Wissenschaften, Hertie AI und des Sonderforschungsbereichs „Robust Vision“ unterstützen.

Vom Anbieter wird erwartet, dass er ein bindendes Angebot, passend zu den folgenden genannten Rahmenbedingungen, einreicht. Die Ausschreibung umfasst Lieferung, Aufbau, Inbetriebnahme und Abnahme des Gesamtsystems am Höchstleistungsrechenzentrum Stuttgart, (HLRS), sowie Garantie, Reparaturen und Support.

2. Rahmenbedingungen

Das Gesamtsystem gliedert sich in verschiedene Teilkomponenten, die im Folgenden detailliert beschrieben werden. Das Finanzvolumen für das Gesamtsystem beträgt 9.652.660,-€ (inkl. Mwst.) bzw. 8.111.479,-€ (ohne Mwst.)

Ein Überschreiten der zur Verfügung stehenden Summe führt zum Ausschluss des Angebots.

Die Beschaffung erfolgt unter dem Vorbehalt, dass zumindest ein Teil der für die Beschaffung zur Verfügung stehenden Mittel (492.660 Euro (inkl. Mwst)) rechtzeitig zum Jahresende abfließen und bis spätestens 31. Dezember 2026 verausgabt werden kann.

Hierfür ist Voraussetzung, dass die Ausführung der Leistung in enger Abstimmung mit dem Auftraggeber ggfls. auch Teillieferungen in der entsprechenden Auftragshöhe vorsieht und diese Teillieferungen bis spätestens Mitte KW 51 geliefert und in Rechnung gestellt werden können.

Mit Abgabe seines Angebots bestätigt der Bieter, bis spätestens zum hier angegebenen Termin und in enger Abstimmung mit dem Auftraggeber, die geforderte (Teil-)Leistung erbringen zu können. Der Auftraggeber behält sich vor im Fall der Nicht-Einhaltung der vorab genannten Vereinbarung bzw. Lieferfrist vom von dem Finanzierungsvorbehalt betroffenen Teil-Auftrag zurückzutreten und den Gesamt-Auftragswert der Beschaffung entsprechend zu reduzieren.

Im Fall eines solchen (Teil-)Rücktritts aus vorgenanntem Grund können vom Auftragnehmer keine Kosten oder Schadenersatz gegenüber dem Auftraggeber geltend gemacht werden.

Ansprechpartner am Tübingen AI Center der Universität Tübingen ist Herr Dr. Simon Kreuzer (simon.kreuzer@uni-tuebingen.de).



2.1. Infrastruktur und Betrieb

Die Hardwarekomponenten werden am HLRS (Höchstleistungsrechenzentrum Stuttgart) installiert (Lieferadresse: Nobelstraße 19, 70569 Stuttgart, Deutschland)

und von den Mitarbeitern des Tübingen AI Centers betreut.

Die Stromversorgung, Anschluss an das Kühlungssystem und Netzwerkanschluss werden vom HLRS bereitgestellt.

Informationen hierzu befinden sich im Anhang dieses Dokumentes (siehe Kapitel 7).

2.2. Begriffe, Abkürzungen und Definitionen

Um Mehrdeutigkeiten zu verhindern, sind die in den von den Anbietern eingereichten Antworten, Konzepten und Angeboten verwendeten Begriffe und Abkürzungen entsprechend den folgenden Definitionen zu verwenden.

2.2.1. *Definition: Gesamtsystem/Bestandsystem*

Das Gesamtsystem umfasst sämtliche Hard- und Softwarebestandteile dieser Ausschreibung und das Bestandsystem. Es beinhaltet sämtliche in diesem Dokument beschriebenen Rechnertypen, Chassis oder Bladecenter, Netzwerkkomponenten inklusive der notwendigen Kabelverbindungen sowie Plattenspeicher. Das Bestandsystem umfasst sämtliche Hard- und Softwarebestandteile sowie Infrastruktur, die bereits am HLRS verbaut wurden.

2.2.2. *Definition: Server*

Ein Server bezeichnet eine Hardwareeinheit mit einer oder mehreren CPUs, eigenem Arbeitsspeicher, eigener Systemfestplatte, eigenen Netzwerkkarten in einem eigenen Gehäuse. Mehrere Gehäuse können unter Umständen in einem Chassis zusammengefasst werden.

2.2.3. *Definition: Storage*

Ein Storage bezeichnet eine Einheit aus einem oder mehreren Speichersystemen und, wenn notwendig, einem oder mehreren Servern.

2.2.4. *Definition: Speichermedien*

Der Ausdruck Speichermedien wird als Oberbegriff für alle Arten von Festplatten oder ähnlichem verwendet und umfasst sowohl Platten mit drehenden Komponenten als auch SSDs oder NVMe. Im Zweifel ist der genaue Typ immer präzise zu spezifizieren.

2.2.5. *Definition: Reparaturfall*

Reparaturen sind auszuführen, wenn Meldungen in Logs auf CPU-, Cache-, Arbeitsspeicher- oder Plattenausfälle oder vergleichbare Störungen hinweisen oder die Homogenität des Systems nicht mehr gegeben ist, das System instabil ist oder überhaupt nicht mehr arbeitet.



2.2.6. Definition: CPU-Kerne (Anzahl CPU-Kerne)

Bei der Anzahl von CPU-Kernen pro Server wird von real existierenden, physischen Kernen ausgegangen und nicht von der virtuell möglichen Anzahl an Kernen nach der Berücksichtigung von Hyperthreading bzw. Simultaneous-Multithreading.

2.2.7. Definition: Byte Größen

Im Kontext dieser Ausschreibung gelten folgende Größen:

Megabyte 1.000.000 (Millionen) Byte

Gigabyte 1.000.000.000 (Milliarde) Byte

Terabyte 1.000.000.000.000 (Billionen) Byte

3. Bewertungskriterien

3.1. Anforderungen der Kategorie A, Ausschlusskriterien (K.o.-Kriterien)

Alle Anforderungen, die als Ausschlusskriterien (A) gekennzeichnet sind, sind obligatorisch und müssen von dem Anbieter erfüllt werden. Angebote, die ein oder mehrere Ausschlusskriterien nicht erfüllen, werden nicht akzeptiert und das Angebot vom Verfahren ausgeschlossen.

Alle Ausschlusskriterien sind Mindestanforderungen, die übertroffen werden können.

Dieses führt jedoch nur dann zu einer besseren Bewertung, wenn die Anforderung zusätzlich auch explizit als Bewertungskriterium (B) aufgeführt ist.

3.2. Anforderungen der Kategorie B, Bewertungskriterien

Alle mit (B) gekennzeichneten Anforderungen sind Bewertungskriterien, deren jeweiliger Erfüllungsgrad durch das Angebot mit Punkten bewertet wird. Die Nichterfüllung eines solchen Kriteriums führt also nicht zum Ausschluss des Angebots, sondern zu einer Bewertung des jeweiligen Kriteriums mit null Punkten.

4. Ortsbesichtigung (verpflichtend):

Für die Erstellung eines aussagekräftigen Angebots wird die Teilnahme an einer Ortsbesichtigung am HLRS Stuttgart für zwingend erforderlich erachtet.

Die Teilnahme an einer Ortsbesichtigung bildet eine zwingende Voraussetzung für die Zulassung eines zum Vergabeverfahren eingereichten Angebots.

Für die Durchführung der Ortsbesichtigung ist der Zeitraum vom 07. September 2026 – 25. September 2026 vorgesehen. Für die Teilnahme ist (schriftlich) ein Termin zu vereinbaren.



Ansprechpartner für die Terminvereinbarung ist:

Simon Kreuzer

Tel. 07071/29-70863

E-Mail: simon.kreuzer@uni-tuebingen.de

Die Frist für die Terminvereinbarung beginnt mit der Eröffnung des Vergabeverfahrens und endet spätestens am 21. September 2026 (12:00 Uhr).

Anmeldungen die außerhalb der genannten Fristen eingehen, können leider nicht berücksichtigt werden. Zwischen Anmeldung und Besichtigungstermin ist ein Vorlauf von wenigstens 3 Arbeitstagen einzukalkulieren.

Im Zuge der Ortsbesichtigung erhalten Sie einen entsprechenden Nachweis der Teilnahme (Vordruck: vgl. Anlage 6).

Der Vordruck ist vorausgefüllt mit den Bieterangaben zum Termin mitzubringen und vom berechtigten Vertreter des Auftraggebers gegenzeichnen zu lassen.

Der vom Auftraggeber unterzeichnete Nachweis ist dem Angebot beizufügen.

Ohne Ortsbesichtigung und entsprechend mit dem Angebot eingereichter Bestätigung kann ein Angebot nicht gewertet werden und wird aus dem Verfahren ausgeschlossen.

Eine Kostenerstattung für die Teilnahme an der Ortsbesichtigung ist nicht vorgesehen.

Wir bitten um Beachtung, dass während der Ortsbesichtigung keine Bieterfragen beantwortet werden. Evtl. sich dort ergebende Bieterfragen sind schriftlich, im Anschluss an die Besichtigung, über das Vergabeportal zu stellen und werden darüber beantwortet.

5. Technische Spezifikationen und Konfiguration

5.1. Übersicht Bestandsystem

Die aktuelle Installation des Conway Clusters ist in zwei Racks untergebracht. In dem einen Rack befinden sich eine In-Rack-CDU und die GPU-Rechenknoten sowie drei Switches. Im zweiten Rack sind Login- und Hypervisor-Knoten sowie die All-Flash-Knoten des Speichers installiert, ebenso drei Switches.

Die Rechenknoten im ersten Rack sind direkt flüssigkeitsgekühlt. Das Rack ist deshalb auf der Rückseite mit Steigleitungen ausgestattet, die Kühlflüssigkeit zwischen der CDU und den Rechenknoten transportiert.

Alle Systeme im zweiten Rack sind primär luftgekühlt.



Zwischen den Racks befindet sich ein Side Cooler, der zusammen mit den Racks eine abgeschlossene Umgebung bildet und die warme Luft aus beiden Racks in Wasser überführt.

Es gibt ein dediziertes Management-Netzwerk auf Basis von Gigabit-Ethernet, das alle Management-Ports der Systeme (Switches, Side Cooler, CDU, BMCs der Knoten) verbindet. Ein weiteres Ethernet (10 Gbit/s) dient der Provisionierung und Administration sowie dem (externen) Datenverkehr.

Das Hochgeschwindigkeits-Netzwerk ist eine NDR-InfiniBand-Fabric, die alle InfiniBand-Adapter der Knoten miteinander verbindet. Derzeit sind alle HCAs mit einem einzigen Switch verbunden (Stern-Topologie).

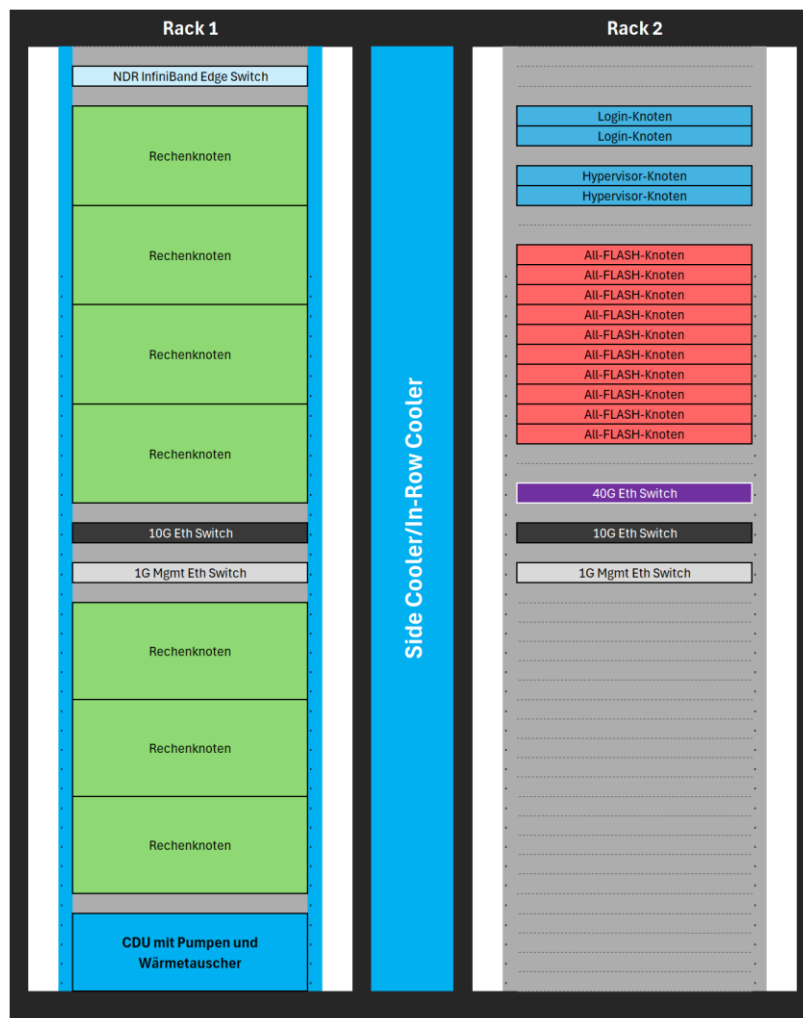


Abb. 1: Schematische Darstellung des Bestandssystems

5.2. Allgemeine Anforderungen

Bitte beachten Sie, dass diese allgemeinen Anforderungen übergreifend für alle nachfolgenden



Positionen der Ausschreibung gelten.

Req. No	Kategorie	Beschreibung
AA-1	A	Nach Erteilung des Zuschlages ist der Rackplan mit dem Auftraggeber abzustimmen. Rechtzeitig vor der Rechnerinstallation sind vom Auftragnehmer alle für die Installation notwendige Angaben in deutscher Sprache zu übergeben. Es ist zeitnah nach dem Zuschlag ein vermaßter Bodenausschnittsplan in M 1:50 oder 1:20 als dwg- und pdf-Datei zu erstellen und die praktische Umsetzbarkeit der Installation vor Ort mit der HLRS Betriebstechnik abzustimmen.
AA-2	A	Der Airflow und der Kühlwasserfluss aller Komponenten ist mit dem Auftraggeber abzustimmen.
AA-3	A	Alle angebotenen Knoten müssen über redundante Netzteile, inklusive Stromkabel zur Verbindung mit C14 oder C19 Stromdosen verfügen, welche an getrennte Stromkreise angeschlossen werden.
AA-4	A	Strom- und Netzkabel sind an beiden Enden mit Labeln zu versehen. Netzkabel müssen mit der Bezeichnung beider Endpunkte beschriftet werden. Stromkabel müssen mit der Bezeichnung des Endpunktes und einer fortlaufenden Nummerierung beschriftet werden. Die entsprechenden Bezeichnungen sind mit dem Auftraggeber nach Erteilung des Zuschlages abzustimmen.
AA-5	A	RJ45 – Kabel sind vom Typ CAT 6A und farbcodiert: Management (IPMI) rot, Provisionierung blau, 1 G beziehungsweise 10 G Ethernet-Switch Uplink-Kabel gelb, Infrastruktur-Telemetrie grün.
AA-6	A	Es ist mindestens eine Herstellergarantie sowie Lizenz- und Gewährleistung von 60 Monaten für alle Knoten, Switches und anderen Komponenten anzubieten. Innerhalb dieses Zeitraums müssen für alle Komponenten der Lösung sämtliche Garantie, Gewährleistungen und Support zur Verfügung stehen. Beginn des Garantie-, Gewährleistungs- und Support-Zeitraums ist der Tag der Endabnahme. Er endet mit Ablauf der im Vertrag stehenden Frist. Mit schriftlicher Mängelrüge wird die Verjährungsfrist unterbrochen und wird mit Behebung der Mängel fortgeführt. Darüber hinaus angebotene Gewährleistungs- und Garantie-Bedingungen sind im Zuge der Angebotsabgabe zu beschreiben.
AA-7	A	Der Anbieter/Auftragnehmer gewährleistet eine Nachlieferungs- und Verfügbarkeitsgarantie für alle Regel- und Ersatzteile von mind. 5 Jahren (ggfls. inkl. Software-Updates).



AA-8	A	Der Anbieter/Auftragnehmer stellt Sicherheitsupdates und kritische Updates für alle Softwarekomponenten innerhalb von einer Woche nach deren Veröffentlichung bereit.
AA-9	A	Updates für alle gelieferten Software-Produkte sind mindestens alle 6 Monate vorzusehen. Der Anbieter/Auftragnehmer stimmt sich mit dem ML Cloud Team ab, um ein entsprechendes Zeitfenster für nicht-kritische Software-Updates zu schaffen.
AA-10	A	Die Abwicklung von Garantieleistungen erfolgt für den Auftraggeber ausschließlich über den Auftragnehmer, sofern nichts anderes vereinbart ist. Im Rahmen der Angebotsabgabe, bzw. spätestens nach Zuschlagserteilung, sind dem Auftraggeber entsprechende Kontaktdaten und Ansprechpartner mitzuteilen.
AA-11	A	Störungsmeldungen erfolgen in elektronischer Form. Der Auftragnehmer stellt dazu entweder eine Emailadresse oder einen Zugang zu einem vom Auftragnehmer betriebenen Ticketsystem zur Verfügung.
AA-12	A	Für alle Software- und Hardwarekomponenten gelten folgende Service Zeiten: Montag bis Freitag 08:00 – 18:00. Für Severity 1 Tickets (mindestens 35% aller Rechenknoten nicht funktionstüchtig, Performance des Netzwerks oder der Rechenknoten um mindestens 35% verringert, CVE mit lokaler Rechteausweitung (LPE)) wird eine Reaktionszeit von höchstens 120 Minuten gefordert, für alle anderen Fälle wird eine Reaktionszeit von maximal vier Stunden gefordert. Als Reaktionszeit ist hier die Zeit gemeint, innerhalb derer bei Gewährleistungs- oder Support-Fragen eine Kontaktherstellung sowie der Beginn der Analyse des Problems durch den Anbieter erwartet wird.
AA-13	A	Der Auftragnehmer stellt sicher, dass die Kontaktaufnahme für den Auftraggeber, hinsichtlich der Abwicklung von Garantie- und Supportleistungen, zu den üblichen Geschäftszeiten kostenfrei und über deutsch- sowie englischsprachiges Personal erfolgen kann.
AA-14	A	Für entfernten Supportzugriff auf das System z.B. über SSH muss 2-Faktor Authentifizierung verwendet werden. Hierfür stellt das ML Cloud Team gegebenenfalls Hardwaretokens zur Verfügung.
AA-15	A	Der Anbieter fungiert als „single point of contact“ (SPOC), d.h. der gesamte Support und anderweitiger Kontakt besteht ausschließlich zu der anbietenden Firma. Reparaturen und Service werden ausschließlich durch Angestellte des Anbieters durchgeführt.



AA-16	A	Insbesondere muss über die gesamte Gewährleistungslaufzeit dem Projekt ein technischer Ansprechpartner fest zugeordnet werden, der mit dem System gut vertraut ist. Der Ansprechpartner muss telefonisch zu den üblichen Geschäftszeiten erreichbar sein und sich persönlich um den Support kümmern. Der Ansprechpartner muss in der Lage sein, technische Probleme des Clusters zu analysieren und in vielen Fällen selbst zu beheben.
AA-17	A	Für alle Hardwarekomponenten gilt ein Next-Business-Day Service. Es dürfen im Rahmen von Supportfällen keine Kosten (z.B. Versandkosten) für die Universität Tübingen entstehen. Field replaceable Units werden durch den Anbieter ersetzt.
AA-18	A	Ersatzteile für customer replaceable units müssen vorgehalten werden. Insbesondere muss von jedem Switch-Typ ein Exemplar vorgehalten werden.
AA-19	A	Die Kosten für Service- und Wartungsarbeiten nach Ablauf der Garantie sind in einer Anlage beizufügen. Die Wartung muss einen kompletten Funktionstest beinhalten und in einem Wartungsprotokoll dokumentiert werden.
AA-20	A	Der Anbieter legt ein Servicekonzept vor, in dem die zur Erfüllung der Wartungsanforderung, Wartungsintervalle und der benötigten Ersatz- und Verschleißteile, sowie Folgekosten dargestellt sind.
AA-21	A	Der Support und etwaige Lizenzen müssen vom Anbieter für den Gewährleistungszeitraum von 60 Monaten zwingend angeboten werden.
AA-22	A	Der Anbieter bietet Vorabaustausch von austauschbaren Einzelkomponenten an, d.h. Ersatzteile werden versendet bevor die mangelhafte/n oder geschädigte/n Komponente/n an den Anbieter überbracht wurde/n.
AA-23	A	Es handelt sich um ein „turnkey-Cluster“, bei dem der Anbieter das gesamte System aufbaut und in Abstimmung mit dem Auftraggeber konfiguriert. Insbesondere muss eine Integration in das bestehende User Management und das job scheduling system SLURM (unter Berücksichtigung der Isolation von nicht exklusiven Jobs (cgroups), die Zuordnung von SMT-Threads und die Zuordnung von GPUs zu physikalischen CPU-Cores) sowie eine Lösung zum automatisierten Aufspielen von Patches und Updates auf die Rechenknoten implementiert werden. Das Provisionieren der Computenodes mit OS-Images muss über eine eigene Lösung, die im bestehenden Proxmox Cluster als VM betrieben wird, realisiert werden. Alternativ können



		die neuen Computenodes auch in das bestehende Provisionierungssystem integriert werden.
AA-24	A	Alle in Servern verbauten SSDs oder NVMe-SSDs unterstützen „power loss protection“.
AA-25	A	Sofern nicht anders gefordert, sind alle Komponenten mit redundanter Stromversorgung ausgestattet.
AA-26	A	Der Auftraggeber legt großen Wert auf einen energieeffizienten Betrieb aller Rechner- und Speicherressourcen. Alle nach Stand der Technik üblichen Stromsparmechanismen der Komponenten (wie beispielsweise Schlafzustände von Prozessoren) müssen vom Auftragnehmer aktiviert werden.
AA-27	A	Komponenten bzw. Funktionalitäten, deren Umfang und Eigenschaften abgefragt werden, sind - vollständig - in die ausführliche technische Spezifikation aufzunehmen.
AA-28	A	Alle angebotenen Systeme inklusive aller Komponenten sind für einen Serverbetrieb von 365 x 24 Stunden pro Jahr ausgelegt.
AA-29	A	Alle Systeme verfügen über einen dedizierten Baseboard Management Controller (BMC) mit eigener, physisch getrennter Management-Netzwerkschnittstelle für Out-of-Band-Remote-Management. Der BMC unterstützt die Redfish-API (DMTF) sowie IPMI 2.0 oder neuer. Folgende Funktionen müssen ohne Einschränkung nutzbar sein: • Remote-KVM / grafische HTML5-Konsole (Console Redirection) • Virtual Media (Einbindung entfernter ISO-/Image-Dateien) • vollständiger, schreibender Zugriff über die Redfish-API (nicht auf Read-only beschränkt) • Out-of-Band-Firmware-Update sowie Telemetrie/Sensor-Auslesung (Temperatur, Leistungsaufnahme) Sämtliche hierfür erforderlichen Lizenzen sind unbefristet und ohne wiederkehrende Kosten mitzuliefern. Die konkrete Lizenzstufe je System ist im Angebot auszuweisen.
AA-30	A	Die Kabellängen müssen passend zu den Distanzen gewählt werden und vor der Installation mit dem Auftraggeber abgestimmt werden. Für das Infiniband Netzwerk müssen ab mehr als 2 m Kabellänge optische Kabel genutzt werden.



AA-31	A	Alle zur Verbindung der Komponenten benötigten Kabel, einschließlich der Netzwerk- und Stromanschlusskabel, sind mitzuliefern.
AA-32	A	Der Auftragnehmer muss sicherstellen, dass die Server über den Up-link des Systems in das dedizierte akademischen Netzwerk BelWü erreichbar sind.
AA-33	A	Der Auftragnehmer muss vor der Abnahme PDF-Dateien aller Hardware- und lizenzierten Software-Handbücher zur Verfügung stellen. Insbesondere muss eine Dokumentation für das Clustermanagementsystem zur Verfügung gestellt werden.
AA-34	A	Der Auftragnehmer muss Zugriff via Root-Passwort auf alle Komponenten zur Verfügung stellen, einschließlich aller erforderlichen Zugriffsverfahren.
AA-35	A	Die Firmware- und BIOS-Stände aller Komponenten werden vom Anbieter geprüft und bis zur Betriebsbereitschaftserklärung auf die letzte, als stabil bezeichnete und für alle Komponenten kompatible Version gebracht.
AA-36	A	Ausführliche Einweisungen und Schulungen der Administratoren des Auftraggebers in angemessenem Umfang sind mit anzubieten und einzupreisen.
AA-37	A	Die Leistungsaufnahme aller Hardwarekomponenten muss einzeln auslesbar sein.
AA-38	A	Als Abnahmetests sind die folgenden Benchmark-Läufe vom Auftragnehmer zu demonstrieren und die Schritte zur Durchführung für den Auftraggeber zu dokumentieren, so dass er sie bei Bedarf in gleicher Weise wiederholen kann. Für die Durchführung der Abnahmetests muss SLURM zusammen mit Singularity/Apptainer als Ausführungsumgebung genutzt werden. Alle Installations- und Konfigurationschritte für diese Benchmarks gehören zum Leistungsumfang. Diese Tests dienen dem Nachweis eines stabilen Betriebs aller Komponenten und einer ausreichenden Kühlung unter Last. - Abnahmetest 1 - HPL: Dieser Test besteht aus dem HPL (High Performance Linpack) laut https://top500.org/project/linpack/ und http://www.netlib.org/benchmark/hpl/. Er ist in einer Weise auszuführen, dass alle Knoten beteiligt werden und alle GPUs genutzt werden. Es ist mindestens eine Effizienz von 40% zu erreichen (Rmax/Rpeak). - Abnahmetest 2 - Single-Node DNN-Training mit syntheti-



		schen Daten: Der auf PyTorch basierende Benchmark wie unter https://github.com/horovod/horovod/blob/master/examples/pytorch/pytorch_synthetic_benchmark.py veröffentlicht ist. Dabei müssen alle GPUs pro Knoten verwendet werden. Dieser Benchmark soll auf allen Knoten gleichzeitig (trivial parallel) ausgeführt werden. Es sind ausschließlich Open-Source Pakete zu verwenden, mit mindestens PyTorch 1.12.1 sowie Horovod v0.26.1 oder neuer. • Abnahmetest 3 - Single-Node Training mit realen Daten: Es ist der PyTorch-basierende Benchmark wie unter https://github.com/horovod/horovod/blob/master/examples/pytorch/pytorch_imagenet_resnet50.py zu verwenden, jedoch mit realen Trainingsdaten, die vom ALL-FLASH Speicher gelesen werden. Es ist der ImageNet-Datensatz ILSVRC2012 zu verwenden. • Abnahmetest 4 - Multi-Node DNN-Training mit synthetischen Daten: Der auf PyTorch basierende Benchmark wie unter https://github.com/horovod/horovod/blob/master/examples/pytorch/pytorch_synthetic_benchmark.py veröffentlicht ist, um auf allen Knoten und GPUs einen Multi-Node-Trainingslauf durchzuführen. Es sind ausschließlich Open-Source Pakete zu verwenden, mit mindestens PyTorch 1.12.1 sowie Horovod v0.26.1 oder neuer. • Abnahmetest 5 - Multi-Node DNN-Training mit realen Daten: Ähnlich zu Abnahmetest 3 ist der auf PyTorch basierende Benchmark https://github.com/horovod/horovod/blob/master/examples/pytorch/pytorch_imagenet_resnet50.py zu nutzen, um einen ähnlichen Multi-Node-Trainingslauf wie unter Test 4 durchzuführen, der reale Daten verwendet. Es soll der ImageNet-Datensatz ILSVRC2012 benutzt werden, der auf dem ALL-FLASH Speicher bereitgestellt wird. Abgesehen davon gelten die gleichen Bedingungen wie für Abnahmetest 4. • Alle fünf Abnahmetests müssen jeweils mindestens 12 Stunden ohne Unterbrechungen auf allen Knoten des Systems laufen. Es dürfen währenddessen keine Ausfälle oder signifikante Verminderungen der Rechenleistung wegen Überhitzung von Komponenten auftreten.
AA-39	<i>B</i>	Die Reaktionszeit auf Severity 1 Tickets wird bewertet



AA-40	<i>B</i>	Die Reaktionszeit auf allgemeine Supportfälle wird bewertet
<i>Tabelle 1: Allgemeine Anforderungen</i>		

Wie im Leistungsverzeichnis beschrieben, soll die Systemerweiterung im Serverraum des HLRS am Standort "Nobelstraße 19, 70569" in Stuttgart installiert werden. Die Vorgaben des HLRS zur Infrastruktur (Strom, Kühlung, Netzwerk) in 7.1 sind unbedingt zu beachten.

5.3. Netzwerke

Die Erweiterung soll analog zum bestehenden System mit drei Netzwerken ausgestattet werden, darunter ein Management-Netzwerk (1 Gbit/s Ethernet), ein Provisionierungsnetz (10 Gbit/s Ethernet) und ein Hochgeschwindigkeitsnetz.

5.3.1. Management-Netzwerk

Das Managementnetz dient zur separaten Anbindung der Managementports aller Clusterkomponenten. Dazu gehören die Managementports der Switch- und Storagekomponenten, die Smart-PDUs, CDUs, die BMCs der Rechenknoten, sowie auch die der Storage- und Managementmaschinen. Alle Komponenten erhalten einen dedizierten Anschluss (im Falle der Rechner-, Login-, Management- und Storageknoten durch einen dedizierten IPMI-Anschluss, „dedicated IPMI“). Dafür müssen gesonderte 1 G Switches genutzt werden („dedicated IPMI“).

Es wird erwartet, dass in jedem Rack ein Management-Switch installiert wird, um Verkabelung über Rackgrenzen hinweg möglichst zu vermeiden. Im bestehenden BMC Netzwerk werden zwei Switches des Typs FS S3900-48T4S_P2C verwendet.

Req. No	Kategorie	Beschreibung
<i>IPMI-1</i>	<i>A</i>	Die BMCs aller Komponenten müssen über dedizierte IPMI-Ports an den Knoten selbst an die dedizierten IPMI-1G-Switches angeschlossen werden.
<i>IPMI-2</i>	<i>A</i>	Jeder 1 G Switch muss redundant mithilfe von 10G Uplinks in das Netzwerk eingebunden sein.



IPMI-3	A	Jeder 1 G Switch verfügt über redundante Netzteile
IPMI-4	A	Jeder Switch muss mindestens 802.3ad Link Aggregation und LACP sowie 4096 VLANs unterstützen (Layer 2).
IPMI-5	A	Mitzuliefern sind alle notwendigen Kabel mit mindestens dem CAT 6A Standard. Alle Kabel müssen an beiden Enden mit den beiden Maschinen beschriftet sein, die sie verbinden.
IPMI-6	A	Es müssen mindestens 40 Ports für zukünftige Erweiterungen zur Verfügung stehen.
Tabelle 2: Anforderungen an das Managementnetzwerk		

5.3.2. Provisionierungsnetz

Das aktuelle Provisionierungsnetz sieht schematisch so aus:

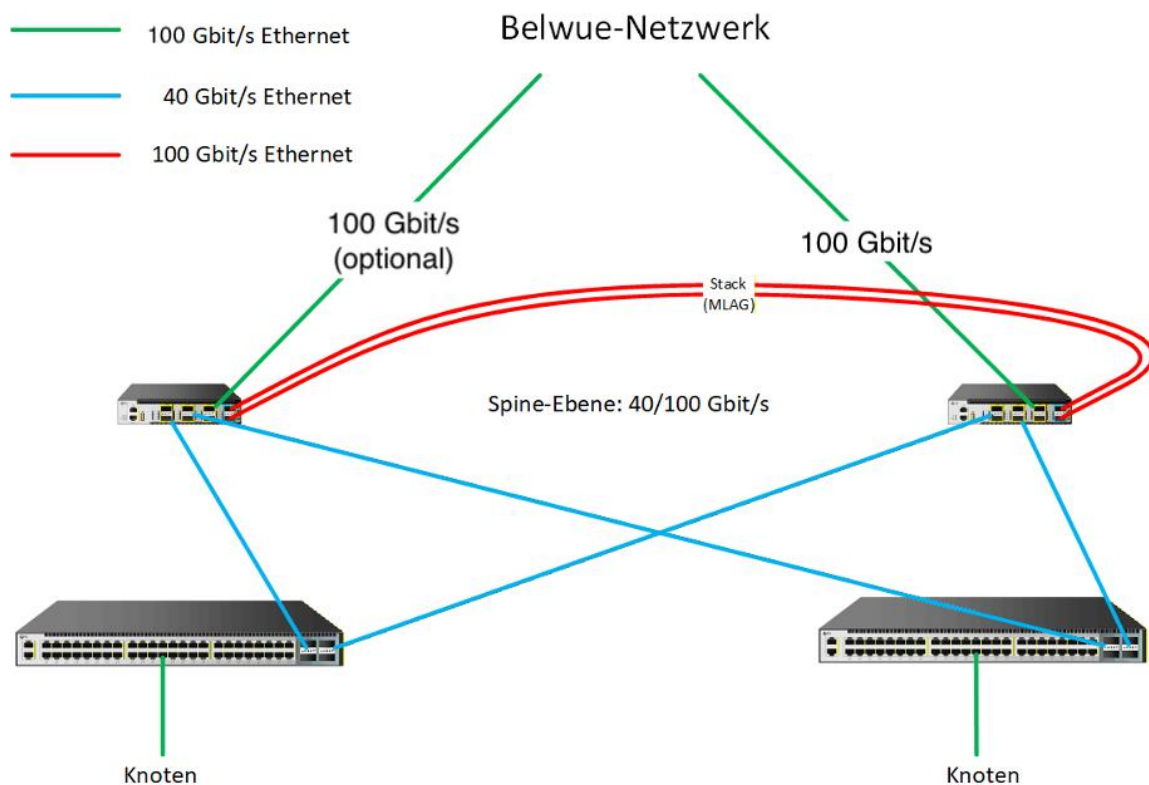


Abb. 1: Schematische Darstellung des Provisionierungsnetzwerkes



Für die Erweiterung des Provisionierungsnetzes wird erwartet, dass in jedem Rack ein Provisionierungsswitch installiert wird, um Verkabelung über Rackgrenzen hinweg möglichst zu vermeiden.

Auf der Spine-Ebene sind derzeit zwei Switches des Modells FS S8550-6Q2C in einer redundanten MLAG-Konfiguration im Einsatz, auf der Leaf-Ebene sind zwei Switches des Modells FS S5860-48XMG im Einsatz. Auf den Spine-Switches sind jeweils mindestens zwei QSFP-Ports (40 Gbit/s) frei, zusammen also mindestens vier. Des Weiteren wird das System ab voraussichtlich Oktober 2026 über einen 100Gbit Uplink verfügen, anstelle der derzeitigen 10 Gbit/s. Die Option für einen weiteren Uplink soll vorgehalten werden.

Die neuen Access-Switches sowie die Access-Switches des Management-Netzwerks müssen redundant mit diesem Stack verbunden werden.

Req. No	Kategorie	Beschreibung
P-1	A	Jeder verwendete Switch verfügt über einen dedizierten Management-Port (RJ45), der an das Managementnetzwerk angeschlossen wird
P-2	A	Es wird ein zentraler, redundanter Switch-Stack für die Spine-Ebene oder eine analoge Lösung erwartet (im Folgenden als „Stack“ bezeichnet). Dieser besteht aus einer Erweiterung des bestehenden Spine-Netzwerks.
P-3	A	Das Stacking der Spine-Ebene ist redundant und erfolgt mit einer Bandbreite von mindestens 100 G je physikalischem Link.
P-4	A	Alle weiteren Switches (Leaf-Switches) sind redundant mit mindestens 40 G je physikalischem Link mit diesem Stack verbunden.
P-5	A	Eine der CPU Nodes muss redundant mit 100G an das Provisionierungsnetzwerk angebunden werden. Diese darf auch direkt mit den Spine-Switches verbunden werden.
P-6	A	Jeder Switch muss mindestens 802.3ad Link Aggregation und LACP sowie 4096 VLANs unterstützen (Layer 2).
P-7	A	Alle Rechenknoten, Login-Knoten, die Speichersysteme sowie die Hypervisor müssen an das Provisionierungsnetzwerk



		angeschlossen werden.
P-8	A	Für eine Erweiterung des all-flash Speichersystem müssen mindestens 20 Anschlüsse im bestehenden Rack berücksichtigt werden.
P-9	A	Mitzuliefern sind alle notwendigen Kabel mit mindestens dem CAT 6A Standard. Ebenso sind alle Kabel und Transceiver für Uplinks mitzuliefern. Alle Kabel müssen an beiden Enden mit den beiden Maschinen beschriftet sein, die sie verbinden.
P-10	A	Das Netzwerk muss so konzipiert sein, dass insgesamt mindestens 40 weitere 10 G Ports für künftiger Erweiterungen in gleicher Weise (möglicherweise mithilfe zusätzlicher Leaf Switches) angeschlossen werden können. Ohne Zukauf zusätzlicher Hardware müssen mindestens 10 weitere Rechenknoten angeschlossen werden können.
P-11	A	Jeder Switch verfügt über redundante Netzteile und redundante Lüfter
P-12	A	Es muss in der Erweiterung des Systems Platz für den Anschluss für bis zu zwei 100Gbit Uplinks an das Belwuenetz vorgesehen werden. Die bestehende Spine-Ebene darf komplett ersetzt werden. Sie muss über mindestens 8 freie 100G Ports verfügen.
Tabelle 3: Anforderungen an das Provisionierungsnetz		

5.3.3. InfiniBand Netzwerk

Mit der InfiniBand-Fabric werden die Rechenknoten untereinander mit sehr hoher Bandbreite verbunden. Ebenso sind Compute- und Storage-Teil des Systems breitbandig miteinander vernetzt.

Neben den Rechenknoten und den Knoten des Speichersystems sind auch alle Hypervisor-Knoten und Login-Knoten an die InfiniBand-Fabric angebunden.

Anders als bei den Ethernet-Netzwerken ist es erlaubt, die Knoten über Rackgrenzen hinweg an einen Switch der InfiniBand-Fabric anzubinden.



Req. No	Kategorie	Beschreibung
IB-1	A	Derzeit sind alle HCAs in sternförmiger Topologie an einen einzelnen Switch angebunden. Während der Installation der Erweiterung müssen alle Systeme (die bestehenden wie auch die neuen) in einer gemeinsamen Fabric zusammengeschaltet werden. Dazu ist eine Tree-Topologie einzurichten, die in der Spine-Ebene aus mindestens zwei Switches besteht. Das Infini-band Netzwerk darf bei Verwendung aller Leaf-Switches höchstens einen Blocking Factor 2,05:1 zwischen der Spine und der Leaf-Ebene aufweisen.
IB-2	A	Kabel und Transceiver müssen seitens dem Hersteller NVIDIA offiziell supportet werden („NVIDIA-validiert“). Als Switches sind nur Produkte erlaubt, die exakt kompatibel zum existierenden NDR-InfiniBand QM9790-Switch sind, z.B. NVIDIA QM9700/QM9790. Die Adapter in den Knoten sind entsprechend zu wählen.
IB-3	A	Es sind 40 weitere NDR Leaf Ports für zukünftige Erweiterungen vorgesehen.
IB-4	A	Alle GPU Rechenknoten werden mit jeweils mind. 8× NDR-400G-Ports (realisierbar z. B. über 4× OSFP-Twinport-Cages, je 2× NDR 400G) Verbindungen an das Netzwerk angeschlossen. Alle Infiniband NICs eines GPU Knotens sollen mit dem selben Leaf verbunden werden, um non-blocking Kommunikation zwischen den GPU Knoten eines Leafs zu erlauben. Alle CPU Rechenknoten müssen mit mindestens einem NDR 400G Port an das Netzwerk angeschlossen werden. Es müssen alle am Rechenknoten verfügbaren NICs mit mindestens 400G mit dem Infinibandnetzwerk verbunden sein.
IB-5	A	Ein optimal konfigurierter Subnetmanager ist einzurichten. Ebenfalls ist die Infiniband Topologie in die bestehen Installation des Ressourcenmanagers Slurm einzupflegen.
Tabelle 4: Anforderungen an das Infinibandnetz		



5.4. GPU - Rechenknoten

Req. No	Kategorie	Beschreibung
C-1	A	Es müssen identische Rechenknoten angeboten werden. Eine Mischkonfiguration ist nicht zulässig.
C-2	A	Es müssen mindestens acht den ARechenknoten angeboten werden.
C-3	A	Alle CPUs unterstützen den PCIe 5.0 Standard, es sind mindestens 128 Lanes verfügbar pro CPU.
C-4	A	Es müssen 8 GPUs pro Knoten angeboten werden, in einer GPU-baseboard-module Konfiguration, in einer geschalteten all-to-all-Fabric über alle GPUs eines Knotens.
C-5	A	Alle angebotenen GPUs müssen • mit NVIDIA CUDA v13 kompatibel sein • eine CUDA Compute Capability ≥ 10.3 besitzen • mindestens 280 Gigabyte GPU-RAM (total) besitzen mit einer Speicherbandbreite von 8 TB/s • von einschlägigen ML- Frameworks (mindestens TensorFlow und PyTorch) für das Training von DNNs direkt unterstützt werden • Multi-Instance-GPU capability besitzen
C-6	A	Alle angebotenen GPU-Baseboard-Module müssen mindestens • 14.4 TB/s interne NVLink Bandbreite besitzen • 2.1 TB GPU memory besitzen • FP4 performance von 144 PFLOPS (sparse) • FP8 performance von 72 PFLOPS (sparse) • Gesamtspeicherbandbreite von 64 TB/s
C-7	A	Die Knoten sind ausgestattet mit zwei x86_64-kompatiblen Prozessoren mit jeweils mindestens 96 physikalischen Cores (192 Threads) pro Prozessor. Jeder Prozessor muss mindestens • eine Grundtaktfrequenz von 2,6 GHz besitzen • DDR5-6400 und PCIe Gen5 unterstützen • 12 Speicherkanäle besitzen • und AVX2 fähig sein



C-8	A	Die angebotenen Knoten müssen mit mindestens 3000 GB Arbeitsspeicher (RAM) ausgestattet sein, die mindestens dem DDR5-6400 Standard entsprechen. Hierfür müssen alle vorhandenen Steckplätze und einheitliche Arbeitsspeicher Module (mit gleicher Kapazität) genutzt werden.
C-9	A	Die angebotenen Knoten müssen zwei 10 G Ethernet Netzwerkinterfaces (RJ45) besitzen, über welchen der Server angesprochen werden kann. Des Weiteren muss ein gesondertes Interface vorhanden sein um den BMC anzusprechen sein muss („dedicated IPMI“).
C-10	A	Die angebotenen Knoten müssen mindestens acht Infiniband Netzwerkports besitzen, die jeweils • mindestens 400 Gb/s Bandbreite unterstützen • mindestens einen OSFP Port besitzen • NDR Infiniband (400 Gb/s) unterstützen • RDMA over Converged Ethernet (RoCE) fähig sind • GPUDirect RDMA unterstützen • PCIe Gen 5.0 kompatibel sind und über mindestens 16x PCIe Lanes angeschlossen werden
C-11	A	Die gelieferten Rechenknoten müssen vom Auftragnehmer in Abstimmung mit dem Tübingen AI Center mit Software-Klienten an das vorhandene Lustre ALL-FLASH System angeschlossen werden, sodass das Filesystem über das Infiniband-Netzwerk gemountet werden kann.
C-12	A	Die angebotenen Knoten müssen für Interfaces im Management-, Provisionierungs- sowie Hochgeschwindigkeitsnetz ein entsprechendes Kabel passender Länge aufweisen
C-13	A	Die angebotenen Knoten müssen für das OS über zwei identische NVMe SSD Speicher in einer Raid 1 Konfiguration verfügen. Die SSD Speicher müssen jeweils • PCIe Gen 4.0 oder höher unterstützen und in einem PCIe-Gen-5-Steckplatz lauffähig sein • über mindestens 480 GB Kapazität verfügen • eine Ausdauer von mindestens 0,3 DWPD über 5 Jahre aufweisen



C-14	A	Die angebotenen Knoten müssen zusätzlich über NVMEs mit insgesamt mindesten 7,68 TB scratch storage verfügen. Ebenfalls müssen mindesten 2 Steckplätze für eine mögliche Erweiterung frei bleiben. Das Volumen erreicht aggregiert 30GB/s für sequenzielles Lesen. Die verwendeten NVME Speicher müssen: • PCIe Gen 5.0 nativ unterstützen (mindestens 4 Lanes) • über mindestens 1,92 TB Kapazität verfügen • eine durchgängige sequentielle Mindestlesegeschwindigkeit von 10.000 MB/s sowie eine durchgängige sequentielle Mindestschreibgeschwindigkeit von 4.000 MB/s erreichen • mindestens 300.000 zufällige Lese- und Schreiboperationen (IOPS) in einer 70/30-Mischung erreichen • mindestens 700.000 zufällige Leseoperationen (IOPS) erreichen • mindestens 150.000 zufällige Schreiboperationen (IOPS) erreichen • eine Latenz von maximal 90 Mikrosekunden (lesend) und 20 Mikrosekunden (schreibend) bei QD1 besitzen • eine Ausdauer von mindestens 1 DWPD über 5 Jahre aufweisen
C-15	A	Alle NVMe SSDs müssen jeweils mit mindestens zwei PCIe Gen 5.0 kompatiblen Lanes angeschlossen sein. Die Anzahl der Lanes muss so gewählt sein, dass die volle Lese- und Schreibgeschwindigkeit der SSDs abgerufen werden kann.
C-16	A	Der Ausfall mehrerer Netzteile kann toleriert werden (N+N Redundanz).
C-17	A	Die Knoten dürfen maximal acht Höheneinheiten im Rack einnehmen. Alle Komponenten müssen mit der maximal möglichen und vom Hersteller vorgesehenen Leistungsaufnahme betrieben werden. Eine Begrenzung der Leistungsaufnahme, insbesondere der GPUs, muss möglich sein, darf aber nicht als Teil der Lösung vorgesehen werden.
C-18	A	Es muss eine Direct-Chip-Cooling Lösung (DLC) verwendet werden.



C-19	A	Alle Knoten müssen gleichzeitig 60h unter maximaler Last auf allen CPUs und GPUs ohne Leistungseinbußen betrieben werden können. Dies muss durch die Abnahmetests 1-5 belegt werden.
C-20	A	Jeder Knoten bietet eine KI-Leistung von mindestens 17.6 PetaFLOPS (theoretische Peak- Performance mit TF32 Datentyp, sparse).
C-21	A	Jeder Knoten bietet eine KI-Leistung von mindestens 36 PetaFLOPS (theoretische Peak-Performance mit FP16 Datentyp, sparse).
C-22	A	Jeder Knoten bietet eine KI-Leistung von mindestens 72 PetaFLOPS (theoretische Peak-Performance mit FP8 Datentyp, sparse).
C-23	A	Jeder Knoten bietet eine KI-Leistung von mindestens 144 PetaFLOPS (theoretische Peak-Performance mit FP4 Datentyp, sparse).
C-24	A	Jeder Knoten unterstützt TF32, BF16, FP16, FP8, FP4, und INT8 Datentypen.
C-25	B	Die Anzahl an Rechenknoten wird bewertet.
Tabelle 5: Anforderungen an die Rechenknoten		

5.5. CPU - Rechenknoten

Für Pre- und Postprocessing sollen CPU-Rechenknoten Teil der Erweiterung sein.

Req. No	Kategorie	Beschreibung
CPU-1	A	Die Knoten sind ausgestattet mit zwei x86_64-kompatiblen Prozessoren mit jeweils mindestens 96 physikalischen Cores (192 Threads) pro Prozessor. Jeder Prozessor muss mindestens - eine Grundtaktfrequenz von 2,6 GHz besitzen 12 Speicherkanäle besitzen



		• und AVX2 fähig sein
CPU-2	A	Die angebotenen Knoten müssen mit mindestens 6 GB Arbeitsspeicher (RAM) pro angebotenem CPU Kern ausgestattet sein, mindestens also 1152 GB. Die verwendeten DIMMs müssen mindestens dem DDR5-6000 Standard entsprechen. Die Speicherbestückung ist so zu wählen, dass die spezifizierte Taktrate erreicht wird.
CPU-3	A	Die angebotenen Knoten müssen zwei 10G Ethernet Netzwerkinterfaces (RJ45) besitzen, über welche der Server angesprochen werden kann. Des Weiteren muss ein gesondertes Interface vorhanden sein um den BMC anzusprechen sein muss („dedicated IPMI“). Eine der CPU Nodes muss direkt mit 2x100G an das Provisionierungsnetzwerk angeschlossen werden.
CPU-4	A	Die angebotenen Knoten müssen mindestens einen Infiniband Netzwerkport besitzen, der • mindestens 400 Gb/s Bandbreite unterstützt • mindestens einen OSFP Port besitzt • NDR Infiniband (400 Gb/s) unterstützen • RDMA over Converged Ethernet (RoCE) fähig sind • GPUDirect RDMA unterstützen • PCIe Gen 5.0 kompatibel sind und über mindestens 16x PCIe Lanes angeschlossen werden
CPU-5	A	Die gelieferten Rechenknoten müssen vom Auftragnehmer in Abstimmung mit dem Tübingen AI Center mit Software-Klienten an das vorhandene Lustre ALL-FLASH System angeschlossen werden, sodass das Filesystem über das Infiniband-Netzwerk gemountet werden kann.
CPU-6	A	Die angebotenen Knoten müssen für Interfaces im Management-, Provisionierungs- sowie Hochgeschwindigkeitsnetz ein entsprechendes Kabel passender Länge aufweisen
CPU-7	A	Die angebotenen Knoten müssen für das OS über zwei identische NVMe SSD Speicher in einer Raid 1 Konfiguration verfügen. Die SSD Speicher müssen jeweils • PCIe Gen 4.0 oder höher unterstützen und in einem



		PCIe-Gen-5-Steckplatz lauffähig sein • über mindestens 480 GB Kapazität verfügen • eine Ausdauer von mindestens 0,3 DWPD über 5 Jahre aufweisen
CPU-8	A	Die angebotenen Knoten müssen zusätzlich über mindestens 3,84 TB nutzbaren scratch Storage in Form von NVMEs verfügen. Ebenfalls müssen mindesten 2 Steckplätze für eine mögliche Erweiterung frei bleiben. Das Volumen erreicht aggregiert 10 GB/s für sequenzielles Lesen. Die verwendeten NVME Speicher müssen: • PCIe Gen 5.0 nativ unterstützen (mindestens 4 Lanes) • über mindestens 1,92 TB Kapazität verfügen • eine durchgängige sequentielle Mindestlesegeschwindigkeit von 7.000 MB/s sowie eine durchgängige sequentielle Mindestschreibgeschwindigkeit von 3.000 MB/s erreichen • mindestens 300.000 zufällige Lese- und Schreiboperationen (IOPS) in einer 70/30-Mischung erreichen • mindestens 700.000 zufällige Leseoperationen (IOPS) erreichen • mindestens 150.000 zufällige Schreiboperationen (IOPS) erreichen • eine Latenz von maximal 90 Mikrosekunden (lesend) und 20 Mikrosekunden (schreibend) bei QD1 besitzen • eine Ausdauer von mindestens 1 DWPD über 5 Jahre aufweisen
CPU-9	A	Der Ausfall eines Netzteils kann toleriert werden.
CPU-10	B	Die Anzahl der angebotenen reinen CPU Knoten wird bewertet. Es müssen mindestens 3 CPU Knoten angeboten werden. Es dürfen höchstens 8 CPU Knoten angeboten werden.
Tabelle 6: Anforderungen an die CPU Rechenknoten		

5.6. Kühlung

Die GPUs in diesem Cluster müssen mit einer Wasserkühlung gekühlt werden, welche



kompatibel mit dem im HLRS vorhandenen System ist. Die Spezifikation der Wasserkühlung befindet sich in Anhang.

Das vorhandene Kühlkonzept sieht vor, dass ein Side Cooler die überschüssige Wärme in den Racks in Wasser überführt. Dazu wurde der Side Cooler jeweils zwischen zwei Racks montiert und ist dadurch in der Lage, beide Racks zu kühlen.

Es wird eine Kühllösung gefordert, die in der Lage ist, sämtliche Wärme, die nicht über direkte Chip-Kühlung abgeführt wird, in Wasser zu überführen.

Im Falle eines Side Coolers soll dieser fest mit den beiden angrenzenden Racks verbunden werden.

Die GPU-Rechenknoten müssen direkt flüssigkeitsgekühlt werden (zumindest CPUs und GPUs in allen Rechenknoten). Hierfür muss eine In-Rack CDU geliefert und an den Gebäudewasserkreislauf angeschlossen werden.

Die CDU muss mit redundanten Pumpen ausgestattet sein und einen Wasser-/ Wasser-Wärmetauscher enthalten. Der Auftragnehmer baut einen Sekundärkreislauf für die Kühlung der Systemkomponenten und stattet die Racks mit entsprechend passenden Steigleitungen (Manifolds) aus.

Alle Anschlüsse für die Verrohrung der Knoten in unserem System sind Universal Quick Disconnect in Größe 06 (UQD06). Alle künftigen Systeme sollen ebenfalls UQD06-Anschlüsse haben. Die neuen Manifolds müssen deshalb kompatibel dazu sein.

Die neue Zelle soll prinzipiell unabhängig von der existierenden Zelle gekühlt werden können. Sollte es Möglichkeiten geben, den neuen Sekundärkreislauf mit dem bestehenden zu verbinden, so dass ein Redundanzgewinn entsteht, bitten wir, dies zu beschreiben.

Req. No	Kategorie	Beschreibung
K-1	A	Das Kühlsystem ist in der Lage, sämtliche Wärme, die nicht über direkte Chip-Kühlung abgeführt wird, in Wasser zu überführen.
K-2	A	Das Kühlsystem muss über eine automatische Abschaltung verfügen, wenn interne oder externe Kühlmittlecks festgestellt werden.
K-3	A	Das Kühlsystem muss über redundante Stromversorgungen und Kühlmittelpumpen verfügen.



K-4	A	Das Kühlsystem muss überwacht werden können.
K-5	A	Das Kühlsystem (CDU und weitere Komponenten) muss innerhalb des Clusters selbst in einem Rack montiert sein.
K-6	A	Das Kühlsystem muss vom Hersteller an das Flüssigkeitskühlsystem des HLRS angeschlossen werden, einschließlich Tests unter maximaler Cluster-Leistungslast.
K-7	A	Das Kühlsystem muss eine „turnkey-solution“ sein, sodass die gesamte Konfiguration, Prüfung und physische Integrität vom Hersteller durchgeführt und garantiert wird und der Hersteller eine quantitativ festgelegte, zeitnahe Entschädigung anbietet, falls das Flüssigkeitskühlsystem ausfällt und dadurch quantifizierbare Schäden am Conway-Cluster oder anderen Systemen innerhalb der Einrichtung entstehen.
K-8	A	Wegen der EMAS-Zertifizierung des HLRS muss ein klimaneutrales Kältemittel (GWP < 1) verwendet werden.
Tabelle 7: Anforderungen an das Kühlsystem		

5.7. Server Rack

Diese Ausschreibung umfasst die Lieferung und Installation der Server-Racks für den Cluster. Die Erweiterung des Conway Clusters soll die angefangene Rack-Reihe möglichst nahtlos fortsetzen. Die bisher installierten Racks sehen wie folgt aus:

- Maße (B x T x H): 800 mm x 1400 mm x 2250 mm (48 Höheneinheiten)
- Feste Verbindung einer Rackseite mit einem Side Cooler (In-Row Cooler)
- Racks sind an Vorder- und Rückseite mit verschließbaren Türen (geschlossen, ohne Perforation) und jeweils einer Seitenwand ausgestattet
- Racks sind mit geeigneten Materialien thermisch isoliert, damit ein möglichst geringer Wärmeeintrag in den umgebenden Raum stattfindet

Req. No	Kategorie	Beschreibung
---------	-----------	--------------



R-1	A	Das Rack-System muss Rackspace und Strom für alle Rechen- und Kühlkomponenten und mindestens zwei weitere Rechenknoten der ausgeschriebenen Leistungsklasse bereitstellen.
R-2	A	Das Rack-System muss die Kühlung aller Rechenkomponenten bei maximaler Leistung unterstützen
R-3	A	Das Rack-System muss den räumlichen Gegebenheiten der HLRS-Einrichtung entsprechen (siehe Abschnitt 7.1)
R-4	A	Das Rack-System darf nicht mehr als 6 Laufmeter einnehmen. Es ist innerhalb dieses Systems Platz für Erweiterungen vorzusehen.
R-5	A	Das Rack-System muss um zusätzliche Racks erweiterbar sein, um eine Erweiterung zu ermöglichen.
R-6	A	Das Rack-System muss Fernsteuerbare Stromverteilungseinheiten (Managed Smart-PDUs) für alle Racks bereitstellen, um alle Rechenkomponenten bei maximaler Leistung ausreichend mit Strom zu versorgen. Diese müssen • das Ein- und Ausschalten einzelner Sockets unterstützen • in ihrer Benutzeroberfläche die Namen der jeweiligen Geräte an den entsprechenden Sockets anzeigen • in der Lage sein, eine vom BMC der Systeme unabhängige Stromverbrauchsmessung eines Servers durchzuführen (Gesamtverbrauchsanzeige über gruppierte Sockets)
R-7	A	Für die Rack-Anordnung müssen Grundrisse vorgelegt werden, die mit den Räumlichkeiten des HLRS kompatibel sein müssen (siehe Abschnitt 7.1).
R-8	A	Die Stromverteilung ist so zu planen, dass beim Ausfall einer Komponente (z.B. Netzteil eines Servers) der Fehler isoliert bzw. eingehegt wird und deshalb keine Fehlerkaskade entsteht. Da gebäudeseitig lediglich eine Stromquelle genutzt wird, sind Redundanzen auf Knotenebene und Verteilungsebene wünschenswert, aber nicht zwingend notwendig. Der Auftraggeber legt jedoch großen Wert auf Fehlerisolierung und erwartet eine plausible Darstellung der Planung zur Stromverteilung mit Ausschluss von Fehlerkaskaden.



R-9	A	Die neuen Racks müssen in der Lage sein 19" Systeme aufzunehmen.
R-10	A	Die Racks verfügen über Temperatur- sowie Luftfeuchtigkeitssensoren an der Vorder- sowie Rückseite in drei unterschiedlichen Höhen (Unten, Mitte und Oben), die über ein Monitoringsystem abfragbar sind.
R-11	B	Die Anzahl der benötigten Racks wird bewertet. Eine kompaktere Bauweise erhält mehr Punkte.
Tabelle 8: Anforderungen an die Racks		

5.8. Facility Access & Service

Das System wird am HLRS eingebaut, betrieben und gewartet. Zugang zum HLRS unterliegt der dortigen Zugangsordnung.

5.9. Clustererweiterung

- Diese Ausschreibung berechtigt auch zum Kauf zusätzlicher Hardware und Software vom Anbieter im Wert von bis zu 30% des ursprünglichen Auftragswerts, um den vom Anbieter angebotenen Conway-Cluster zu erweitern.
- Diese zusätzliche Hardware und Software ist als Erweiterung des in dieser Ausschreibung beschriebenen Systems zu betrachten und muss vollständig mit dem Angebot des Anbieters kompatibel sein.

6. Lieferung und Aufbau

Für die Lieferung, den Aufbau und die Abnahme ist der Zeitraum bis Ende 2026 vorgesehen.

Req. No	Kategorie	Beschreibung
D-1	A	Die jeweiligen Systeme werden durch den/die Anbieter an den Nutzungsort geliefert und aufgebaut. Die Aufstellungsbedingungen sind im Anhang Infrastruktur beschrieben.
D-2	A	Der Kühlung aller Komponenten ist mit dem Auftraggeber abzustimmen. Stromanschlüssen sind auf der Rückseite der Sys-



		teme vorzusehen.
D-3	A	Das System wird durch den Anbieter verkabelt, wobei alle Kabel an beiden Enden zu beschriften und nach Netzwerktyp farblich abzuheben sind. Bei der Verkabelung müssen Wiederöffnungs-Kabelbinder (Klettverschluss) eingesetzt werden.
D-4	A	Weiterhin wird vom Anbieter eine Liste mit allen wesentlichen Systemdaten dem Auftraggeber bei Installation beziehungsweise spätestens bei der Abnahme in elektronischer Form übergeben. Diese enthält unter anderem Seriennummern, Firmware-Version und MAC-Adressen. Der Anbieter spricht nach Auftragserteilung Form und Inhalt der Systemdatenliste mit dem Auftraggeber ab. Es ist ebenso ein Rackplan zu erstellen, dem die Rackbelegung klar wird.
D-5	A	Nach dem Aufbau und der Installation erfolgt ein allgemeiner Funktionstest zusammen mit dem Auftraggeber. Einzelheiten werden hierzu nach Vergabe im Rahmen einer Einführung und Schulung besprochen.
D-6	A	Von dem Anbieter wird auf allen Servern die aktuelle Version von RHEL oder ein Derivat in Absprache mit dem Auftraggeber installiert, sowie das gesamte System konfiguriert. Abweichungen oder ergänzende Komponenten sind mit dem Auftraggeber abzustimmen.
D-7	A	Die neuen Rechenknoten müssen entweder in das bestehende Node-Deploymentsystem integriert werden oder es muss ein alternatives Deploymentsystem, das innerhalb der Proxmox-Umgebung als VM betrieben wird, bereitgestellt werden.
D-8	A	Lieferung und Aufbau muss im Jahr 2026 erfolgen können.
D-9	A	CDU und Side Cooler müssen an den Gebäudewasserkreislauf angeschlossen werden
D-10	A	Installation der Systemerweiterung mit allen Netzwerken und Integration der Erweiterung in das bestehende System, hier



		insbesondere Rekonfiguration der InfiniBand-Fabric
D-11	A	Bei der Integration der Erweiterung ist auf eine möglichst geringe Downtime des Bestandssystems zu achten.
D-12	A	Der Auftragnehmer verpflichtet sich zu wöchentlichen Statusberichten, bis das System abgenommen ist, falls nötig ebenso für ein wöchentliches Meeting/Call.
<i>Tabelle 9: Anforderungen an Lieferung und Aufbau</i>		

6.1. Leistungsumfang, Abnahme und Endkontrolle

Die beauftragte Leistung ist vom Auftragnehmer vollständig zu erfüllen; sie tritt mit der erfolgreichen Funktionsabnahme, die analog den Regularien von §8 BVB-Kauf durchzuführen ist, ein. Nach vollständiger Aufstellung, Montage und Inbetriebnahme erfolgt eine förmliche Abnahme der erbrachten Leistung durch den Auftraggeber oder dessen Beauftragten.

Die Abnahme beinhaltet einen oder mehrere Leistungstests die erfolgreich zu absolvieren sind. Sollte sich zeigen, dass die im Angebot zugesicherten Eigenschaften nicht erfüllt sind, ist der Auftragnehmer innerhalb einer vom Auftraggeber zu setzenden und angemessenen Frist zur Nachbesserung verpflichtet. Die Abnahme wird schriftlich protokolliert.

Zeigen sich Mängel an der gelieferten Sache, obliegt es im Fall – auch nach der Abnahme – dem Auftragnehmer nachzuweisen, dass er die Mängel nicht zu vertreten hat.

Festgestellte Mängel werden dokumentiert; sie sind vom Auftragnehmer unverzüglich zu beheben. Das Beheben der Mängel ist schriftlich zu dokumentieren und muss durch Unterschrift des Auftraggebers oder dessen Beauftragten bestätigt werden.



Req. No	Kategorie	Beschreibung
BE-1	A	Während des Abnahmeverfahrens werden die aufgeführten Abnahmeprüfungen verwendet.
BE-2	A	Die Abnahmeverfahren werden vollständig dokumentiert und gemeinsam vom Lieferanten und dem ML Cloud Team durchgeführt.
<i>Tabelle 10: Anforderungen an Abnahme und Endkontrolle</i>		



7. Anhang

7.1. Technische Anschlussbedingungen für die Erweiterung des „Conway“ Clusters am HLRS

Tabelle 11 listet alle technischen Anschlussbedingungen auf.

Was	Wert	Anmerkung
Doppelboden Lastklasse	5	Höhere Bodenlasten sind mit einem Umbau möglich, die Kosten dafür trägt TÜ
maximale Punktlast	5 kN	Höhere Bodenlasten sind mit einem Umbau möglich, die Kosten dafür trägt TÜ
Anzahl Racks max	10 laufende Meter für Serverschränke und CDUs	
Kühlung	Wasserkühlung	
max. Wärmeabgabe an die Luft	max 10%	Bei Einhaltung des Luft-Wärmebudgets könnten Storage, Login usw luftgekühlt werden
Kühlwasser Vorlauf	14 - 25 °C	variabel. +/- 5 °C Regelschwankung je nach Aussentemperatur
Kühlwasser Spreizung	10 +/- 2 °C	
Min. Druckstufe (Nenndruck)	10 bar	
Druckverlust	1,4 bar	
Flexible Anschlussschläuche	Mindestens 10 bar	Panzerschlauch mit Edeltahlgewebe ummantelt, mit genormten Anschluss- Verbindungen (Schlauchschellen nicht erlaubt)
Kühlwasseranschlüsse	DN 50	Im Doppelboden, vorhandene Anzahl ist ausreichend für einen Anschluss pro Rack
Interner Kühlkreislauf	Keine brennbaren Fluide Keine gesundheitsschädlichen Fluide Gering toxische und schwer entflammbare Wasser-Glykol-Gemische sind erlaubt Keine Fluide mit einem GWP > 1	Sicherheitsdatenblätter für interne Fluide sind mitzuliefern
Luftkühlung	Bei Einhaltung des Luft-Wärmebudgets könnten Storage, Login usw luftgekühlt werden	
Raumluft Temperatur	ca. 22 °C	+/- 3 °C
Maximale Gesamt-Anschlussleistung inkl. Bestandsystem	700 kW	Bestehend aus ca. 620 kW dyn.USV und ca. 80 kW Batt.-USV



Stromversorgung Rechenknoten	dyn. USV, max 10 s	
Anschluss Rechenknoten	125 A CEE-Stecker 5-polig IP67 oder	Im Doppelboden, 5-polig ist unbedingt zu beachten (d.h. nicht 4-polig)
	64 A CEE-Stecker 5-polig mit Umbau	
Stromversorgung Storage, login etc	Batt.-USV, max 15 min	
Anschluss Storage, login, etc	32A CEE-Stecker IP67	Im Doppelboden. Max. 4x32A Anschlüsse
Elektrische Sicherheit	DGUV-3 Prüfprotokoll ist vorzulegen	
Anschluss Datennetz BELWUE	Patchfeld E2000 APC Singlemode	Strecke etwa 50m
Energiedatenerfassung Elektro	Messtechnik muss nachgerüstet werden	Die Kosten trägt Tü. Kältemenge = Elektro
Anlieferung		
Lieferanschrift	Höchstleistungsrechenzentrum der Universität Stuttgart	
	Nobelstraße 19	
	70569 Stuttgart	
Serverraum	1.OG	
Doppelboden	EG, Geschosshöhe, steht außer für Kühlung und Kabelführung nicht zur Verfügung	
Max Tragelast Aufzug	2700 kg	
Türgröße Aufzug	B 2,2 m x H 2,3 m	
Nutzbare Grundfläche der Aufzugkabine	B 2,2 m x T 1,8 m	

Tabelle11: Anschluss- und Anlieferungsbedingungen