Leistungsverzeichnis

NHR GPU-Cluster 2025/2026

Ausfüllhinweise: Sie müssen alle farblich unterlegten, unterstrichenen Felder ausfüllen. Optional können Sie Angaben in Feldern machen, die nur unterstrichen, aber nicht farblich unterlegt sind. Tragen Sie in der Spalte "Mengen- und Preisangaben" alle notwendigen, geforderten Angaben ein (Preise und Kosten jeweils ohne gesetzliche USt.). Ist eine Preiseinheit ungleich 1 vorgegeben (z.B. 1.000), so geben Sie bitte den Preis netto pro Einheit bezogen auf die Preiseinheit an (z.B. 10,00 EUR pro 1.000 Mengeneinheiten). Beziehen Sie in Rahmenvertragspositionen Ihren angebotenen Preis auf die angegebene geschätzte Menge. Geben Sie in der Spalte "Gesamtbetrag netto inkl. Pos.- Nachlass (EUR)" für jede Position den Betrag an, der für die Position aus den Einzelangaben zu kalkulieren ist. Tragen Sie ggf. einen auf Positionsebene gewährten Nachlass ohne Bedingungen im entsprechenden Feld in der Spalte "Mengen- und Preisangaben" ein. Beispiel für eine Position mit angegebener Menge und gefordertem Preis: Die Menge ist mit dem Preis netto pro Einheit in Euro, abzüglich einem evtl. auf Positionsebene gewährten Nachlass ohne Bedingungen, zu multiplizieren.

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.- Nachlass (EUR)
1	Formelle Hinweise Aus gegebenem Anlass weisen wir darauf hin, dass eine Einreichung der elektronischen Angebote per E-Mail nicht zugelassen ist. Auch eine Einreichung des Angebots via Bieternachricht ist unzulässig. Zur Einreichung der Angebote nutzen Sie bitte die Vergabeplattform unter Einsatz des Bietercockpits. Zur Vereinfachung und Erhaltung des Wettbewerbes wird auf das vereinfachte Texterfordernis abgestellt, so dass auf eine Signatur oder echte Unterschrift verzichtet werden kann, wenn aufgrund anderer Umstände feststeht, dass ein Bevollmächtigter die Verantwortung für den Inhalt des Angebotes übernimmt. Aus gegebenem Anlass bitten wir Sie auf die Einreichung eines separaten Angebots aus Ihrem Warenwirtschaftssystem (das Änderungen oder Ergänzungen an den Vertragsunterlagen enthält) zu verzichten. Angebote, die Änderungen oder Ergänzungen an den Vertragsunterlagen enthalten, werden zwangsläufig ausgeschlossen. Für alle verwendeten Typ- und Markenbezeichnungen, die zwecks der technischen Verdeutlichung in der Leistungsbeschreibung aufgeführt sind, gilt der Zusatz "oder gleichwertiger Art". Für geforderte Leistungsparameter gilt der Zusatz "oder besser". Bei Abweichung zur Leistungsbeschreibung ist die technische Gleichwertigkeit schlüssig nachzuweisen. Das Vergaberecht ist derzeit noch nicht gendergerecht formuliert, daher werden hier die Begriffe "Bewerber", "Teilnehmer" und "Bieter" verwendet. Wir richten uns damit dennoch an alle Geschlechteridentitäten.		
2	Wettbewerbsregisterauszug Der Auftraggeber wird ab einer Auftragssumme von mehr als 30.000 € für den Bieter, welcher Zuschlag erhalten soll, zur Bestätigung seiner Erklärungen einen Auszug aus dem Wettbewerbsregister beim Bundeskartellamt anfordern. Ein negativ Eintrag kann zum Ausschluss führen.		

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
3	Allgemeine Aufgabenbeschreibung und Vorbemerkungen Das Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) der TU Dresden betreibt seit mehreren Jahrzehnten Hochleistungsrechner für akademische Nutzer/innen aus Sachsen und ganz Deutschland. Im November 2020 wurde es als eines der HPC-Rechenzentren für das "Nationale Hochleistungsrechnen" (NHR) ausgewählt. Damit ist die Finanzierung für u.a. Hardware-Beschaffungen für 10 Jahre verbunden. Aus dem Hardware-Budget für 2025 und 2026 soll eine Beschaffung für einen GPU-Cluster erfolgen, der insbesondere die Bedarfe für Anwendungen des maschinellen Lernens decken soll. Die Ausschreibung wird in einem Verhandlungsverfahren stattfinden. Darin sind Erstangebote basierend auf der vorliegenden Anforderungsbeschreibung abzugeben. Für eine Reihe von Themen sind Lösungsvorschläge für die Realisierung erbeten bzw. möglichst kosteneffiziente Varianten gewünscht. Für diese sollen die Erstangebote Lösungsvorschläge unterbreiten. Die Erstangebote sind bindend und müssen bereits zuschlagsfähig sein. In den Verhandlungen mit allen ausgewählten Bietern werden die Vor- und Nachteile aller Lösungsvorschläge bewertet und die Resultate in eine finale Formulierung des Leistungsverzeichnisses einfließen, wobei auch die Gewichtung und Punkteskala von B-Kriterien vom AG möglicherweise angepasst werden. Darauf können alle Bieter die finalen Angebote abgeben. Die Termine für die Verhandlung in Präsenz in Dresden werden rechtzeitig über die Vergabeplattform bekannt gegeben. Die optionalen Positionen fließen in die Wertung des wirtschaftlichen Angebots mit ein und müssen von den Bietern angeboten werden. Der Auftraggeber hat das Wahlrecht, diese Optionen zu den vom Bieter genannten Preisen anzunehmen. Bitte beachten Sie, dass für die ausgeschriebene Leistung in Position 1 nur ein Budget in Höhe von 8.200.000,00 Euro (brutto, inkl. MwSt.) zur Verfügung stehen. Unter Berücksichtigung der Anforderungen im Positionstext fordern wir Sie auf, den Leistungsumfang bzw. Ihr Angebot zu optimieren, wobei das zur Verfügung stehende Budget für Position 1 keinesfalls überschritten werden darf. Durch die Optionen behält sich der AG das Recht vor, den GPU-Cluster zu erweitern und dafür zusätzliche Mittel zu verausgaben, die zum jetzigen Zeitpunkt noch nicht zugesagt werden können. Der zu beschaffende GPU-Cluster umfasst GPU-Knoten mit dedizierten Host-CPUs. Es sollen Multi-GPU-Knoten mit 4 bis 8 GPUs je Knoten zum Einsatz kommen. Mit GPU ist in diesem Leistungsverzeichnis die kleinste Hardware-Einheit gemeint, in der GPU-Rechenbeschleuniger von GPU-Herstellern angeboten werden, jedoch keine ggf. vorhandene Unterteilung innerhalb dieser Einheit. Zusätzlich sollen Service-Knoten (CPU-Knoten) in geringerem Umfang sowie Management-Knoten angeboten werden. Ein HPC-Interconnect ist anzubieten. Zusätzlich soll ein Speichersystem bereitgestellt werden, das für alle GPU- und Service- und Management-Knoten sehr schnell mit hoher Bandbreite und geringer Latenz über den HPC-Interconnect zugreifbar ist, jedoch nicht als lokale Storage-Devices in den Knoten selbst.		

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
4	Ortsbegehung Im Rahmen der Angebotserarbeitung besteht die Möglichkeit einer Ortsbegehung. Entsprechende Termine können über die Vergabeplattform evergabe.de rechtzeitig vereinbart werden. Achtung: Termine können NICHT direkt über den Nutzer vereinbart und abgewickelt werden - die Vergabestelle koordiniert diese Aufgaben.		
5	Anforderungen an das Angebot Komponenten bzw. Funktionalitäten, deren Umfang und Eigenschaften abgefragt werden, sind - vollständig - in die ausführliche technische Spezifikation (siehe Fragebögen, Bewertungs- und Ausschlusskriterien) aufzunehmen. Aus gegebenem Anlass bitten wir Sie, auf die Einreichung eines separaten Angebots aus Ihrem Warenwirtschaftssystem (das Änderungen oder Ergänzungen an den Vertragsunterlagen enthält) zu verzichten. Angebote, die Änderungen oder Ergänzungen an den Vertragsunterlagen enthalten, werden zwangsläufig ausgeschlossen. Für alle verwendeten Typ- bzw. Markenbezeichnungen, die ggf. zwecks technischer Verdeutlichung in der Leistungsbeschreibung aufgeführt sind, gilt der Zusatz "oder gleichwertiger Art". Bei Abweichungen zur Leistungsbeschreibung sind die technische Gleichwertigkeit und die Passfähigkeit umfassend und schlüssig nachzuweisen, ggf. auf einer separaten Anlage.		
6	Hinweise zum EVB-IT Systemvertrag Bei der Bearbeitung der Unterlagen sind in Anlehnung an die Modifikationen des Auftraggebers (in blau) die Eintragungen des Bieters (in rot) zur Gewährleistung der Nachvollziehbarkeit farbig hervorzuheben und damit kenntlich zu machen. Wir weisen Sie darauf hin, dass der EVB-IT-Systemvertrag bereits mit dem Angebot vollständig auszufüllen ist und dem Angebot beigefügt werden muss.		
7	Gewährleistung (Verjährung der Mängelansprüche) Es sind mindestens 36 Monate Gewährleistung (Verjährung der Mängelansprüche) anzubieten. Rechte des Auftraggebers bei Mängeln am Gesamtsystem regeln sich gemäß Abschnitt 13 der "Ergänzenden Vertragsbedingungen für das Gesamtsystem - EVB-IT System-AGB" sowie im Besonderen gemäß des Abschnitts 14 des EVB-IT-Systemvertrages.		
8	Basis der Preise Die eingesetzten Preise sind Festpreise und beziehen sich auf die angebotenen Ausführungen der jeweiligen Positionen einschließlich Lieferung, Entladen, Verpackung und Aufstellung frei Verwendungsstelle, Entsorgen von Verpackungsmaterial und Installation der Komponenten soweit dies Bestandteil des Leistungsverzeichnisses ist. Eingeschlossen sind hier alle Kosten für Nebenleistung, etwaige Auslösungs-, Fahrt-, Zehr- und Wegegelder, Lohnzulagen, Über- und Sonntagsstunden, welche aus Gründen, die der Auftragnehmer zu vertreten hat, geleistet werden müssen. Nachforderungen des Bieters wegen gestiegener Kosten sind ausgeschlossen.		

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
9	Abnahme Der Auftragnehmer liefert den Nachweis der Funktionsfähigkeit und Zuverlässigkeit der eingesetzten Hardware gemäß Angebot unter Berücksichtigung des Leistungsverzeichnisses im Rahmen einer Abnahme gemäß der Regelungen des EVB-IT Systemvertrags. Über die erfolgreiche Abnahme, einschließlich der entsprechenden Abnahmeversuche, ist ein Abnahmeprotokoll (siehe Anlage) zu erstellen, das durch den Auftraggeber und den Auftragnehmer zu unterschreiben ist. Alle Abnahmetests und Benchmarks sind in der gleichen Systemkonfiguration (Versionen und Einstellungen aller HW- und SW-Komponenten einschließlich Infrastruktur, BIOS, Betriebssystem, Treiber, Software-Stack, Energieeffizienz-Einstellungen) nachzuweisen, die auch für den Nutzerbetrieb vorgesehen ist. Ausgeschlossen sind davon einzig Einstellungen, die im regulären Nutzerbetrieb von Nutzern selbst eingestellt werden können, wie bspw. Frequenzen von CPU und GPU. In Rücksprache mit dem AG sind Benchmarks zu anderen Zwecken mit dahingehend optimierten Einstellungen möglich (z.B. Green500), diese sind jedoch für die Abnahme nicht zulässig. Der AG weist darauf hin, das Abnahmetests VOR Erklärung der Betriebsbereitschaft durchzuführen sind, wenn im jeweiligen Kriterium gefordert. Spätestens mit Erklärung der Betriebsbereitschaft übermittelt der AN die Werte C_SERV sowie C_STOR im Sinne der Delegierten Verordnung (EU) 2024/1364 für das gesamte Angebot an den AG. Diese Werte werden für die Meldung an das Rechenzentrumsregister (RZReg) der Bundesregierung benötigt.		
10	Ausschluss- und Bewertungskriterien A = Ausschlusskriterium: Die Nichterfüllung einer als Ausschlusskriterium gekennzeichneten Anforderung führt zum Ausschluss des Angebotes. Die Ausschlusskriterien sind KO-Kriterien und werden mit "ja" (=erfüllt) oder "nein" (=nicht erfüllt) bewertet. Ein mit "nein" bewertetes Ausschlusskriterium führt zum zwingenden Ausschluss des Angebots. B = Bewertungskriterium: Die als Bewertungskriterium gekennzeichneten Anforderungen stellen die zu bewertenden Kriterien dar und dienen als Grundlage zur Bestimmung des wirtschaftlichsten Angebotes. Zur Bestimmung des wirtschaftlichsten Angebotes werden die Leistungspunkte (LP) herangezogen.		
11	Lieferung Frei Verwendungsstelle Das Angebot umfasst die Lieferung aller Teile frei Verwendungsstelle inklusive Verpackung, Transportversicherung, Entladen sowie Vertragen/Einbringen. Das Angebot enthält alle für die Installation und Inbetriebnahme des Gesamtsystems notwendigen Betriebs- und Verbrauchsmaterialien. Die Lieferanschrift ist Anlage 04 "Betriebskonzept des Lehmann-Zentrum - Rechenzentrum (LZR) der TU Dresden, Anhang A: Anlieferung und Materialtransport" zu entnehmen.		

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
12	Vergütung und Rechnungslegung Die Rechnungen sind ausschließlich an die zentrale Rechnungsanschrift der TU Dresden unter Angabe des Bestellbelegnummer als Referenz zu übersenden: TU Dresden Zentraler Rechnungseingang 01062 Dresden Alternativ können E-Rechnungen eingereicht werden. Anforderungen an die Rechnungslegung entnehmen Sie bitte der Anlage "Hinweis Anforderung Rechnungsstellung".		
12.1	Zahlungsplan Als Zahlungsmodalitäten werden vereinbart: Anzahlung (Anzahlung gegen Bankbürgschaft): 53,17% des Bestellwertes der Position 1 und 100% des Bestellwerts der Position 3 (bei Inanspruchnahme dieser optionalen Position) nach Eingang der Auftragsbestätigung und Rechnungslegung (gemäß § 17 Ziff. 1, Satz 2 VOL/B), innerhalb von 30 Tagen (unter Berücksichtigung von Skonto, soweit angeboten). Vorauszahlungen bzw. Anzahlungen erfolgen nur nach Vorlage einer zeitlich unbefristet ausgestellten, gültigen Bankbürgschaft eines Kreditinstituts aus einem Mitgliedsstaat der EU. Die Bankbürgschaft wird zurückgegeben, sobald die schriftliche Betriebsbereitschaftserklärung dem Sachgebiet Zentrale Beschaffung und Anlagenbuchhaltung (SG 1.2) vorliegt. Teilzahlung (nach Betriebsbereitschaftserklärung): Y % des Bestellwertes der Positionen 1 nach Betriebsbereitschaftserklärung und Rechnungslegung innerhalb von 30 Tagen, sofern keine anderen Vereinbarungen bzgl. Skonti getroffen sind. Abschlusszahlung: Z % des Bestellwertes der Positionen 1 nach erfolgreicher Abnahme und Rechnungslegung innerhalb von 30 Tagen, sofern keine anderen Vereinbarungen bzgl. Skonti getroffen sind. Bitte beachten Sie, dass für die dritte Teilzahlung mindestens 20 % des Auftragswertes (ohne Option) einzutragen sind. Bitte befüllen Sie den nachfolgenden Fragebogen und geben Sie die gewünschten Prozentsätze für Y und Z an. Abschlagszahlung (Wartung 4. Jahr): Position 2 sowie 4 (falls Option in Anspruch genommen) zum Ende des dritten Betriebsjahres zum Preis wie im Angebot festgelegt. Abschlagszahlung (Wartung 5. Jahr): Position 2 sowie 4 (falls Option in Anspruch genommen) zum Ende des vierten Betriebsjahres zum Preis wie im Angebot festgelegt.		

<table>
<tr><th>Nr.</th><th>Bezeichnung</th><th>Mengen- und Preisangaben</th><th>Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)</th></tr>
<tr><td rowspan="3">F 12.1.1</td><td colspan="2">zu 12.1: Info-Fragebogen</td><td rowspan="3"></td></tr>
<tr><td>Fragetitel</td><td>Antwort</td></tr>
<tr><td>1.1 Prozentsatz für die zweite Teilzahlung (Y):
Wie hoch ist der gewünschte Prozentsatz für die zweite Teilzahlung?</td><td></td></tr>
<tr><td></td><td>1.2 Prozentsatz für die dritte Teilzahlung (Z):
Wie hoch ist der gewünschte Prozentsatz für die dritte Teilzahlung?

Bitte beachten Sie, dass Sie mindestens 20 % für Z (Abschlussrechnung) eintragen.

Soweit Sie keine weitere Teilzahlungen (Y) wünschen, tragen Sie für Z 46,83 % ein.</td><td></td><td></td></tr>
<tr><td>12.2</td><td colspan="2">Zahlungsplan Erweiterung

Teilzahlung (Anzahlung gegen Bankbürgschaft): 100% des Bestellwertes der Position 5, falls diese optionale Position in Anspruch genommen wird, nach Eingang der Auftragsbestätigung und Rechnungslegung (gemäß § 17 Ziff. 1, Satz 2 VOL/B), innerhalb von 30 Tagen (unter Berücksichtigung von Skonto, soweit angeboten). Vorauszahlungen bzw. Anzahlungen erfolgen nur nach Vorlage einer zeitlich unbefristet ausgestellten, gültigen Bankbürgschaft eines Kreditinstituts aus einem Mitgliedsstaat der EU. Die Bankbürgschaft wird zurückgegeben, sobald die schriftliche Betriebsbereitschaftserklärung dem Sachgebiet Zentrale Beschaffung und Anlagenbuchhaltung (SG 1.2) vorliegt.

Abschlagszahlung (Wartung 4. Jahr): Position 6 (falls Option in Anspruch genommen) zum Ende des dritten Betriebsjahres zum Preis wie im Angebot festgelegt.

Abschlagszahlung (Wartung 5. Jahr): Position 6 (falls Option in Anspruch genommen) zum Ende des vierten Betriebsjahres zum Preis wie im Angebot festgelegt.</td><td></td></tr>
</table>

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
1	Hardware, Software, Installation und Wartung Hardware, Software, Installation und Wartung des GPU-Clusters gemäß Kriterienhauptgruppe (KHG) A bis H, inkl. Garantie mit Gewährleistung und Support für 3 Jahre (Laufzeit beginnt nach Abnahme). Eingeschlossen ist die vor-Ort Installation aller Hardware- und Software-Komponenten. Nach Installation erfolgt eine Einweisung in die Geräte und die Software. Die Installation schließt mit der Erklärung der Betriebsbereitschaft und des anschließenden Testzeitraums und Abnahme (unter Nutzung des Abnahmeprotokolls).	**Menge:** 1 Stück **Preiseinheit:** 1 Stück **Nettopreis in Euro** _____ **USt.:** 19 %, falls abweichend _____ % **Nachlass (%)** _____	_____
2	Wartung für das 4. und 5. Jahr des Betriebes Kosten für Wartungs- und Instandhaltung für das 4. und 5. Jahr für das System aus Position Nr. 1. Es ist der Preis pro ein Jahr Wartungsverlängerung anzugeben. Diese Position ist relevant für die Angebotssumme und geht in die Ermittlung des wirtschaftlich günstigsten Angebots in voller Höhe für 2 Jahre ein. Der Einzelpreis umfasst den Preis für ein Jahr. Die Mindestanforderungen (=Ausschlusskriterien) an den Leistungsgegenstand ergeben sich aus der Kriterienhauptgruppe (KHG) A bis H.	**Menge:** 2 Jahr **Preiseinheit:** 1 Jahr **Nettopreis in Euro** _____ **USt.:** 19 %, falls abweichend _____ % **Nachlass (%)** _____	_____

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
3	Optionale Position - **relevant für Angebotssumme** Mehrabnahme zum Zeitpunkt des Zuschlags (GPU-Knoten) Weitere GPU-Knoten in identischer Konfiguration wie unter Position 1 entsprechend KHG A bis H inkl. Garantie mit Gewährleistung und Support für 3 Jahre (Laufzeit beginnt nach Abnahme). Der Auftraggeber kann zum Zeitpunkt des Zuschlags entscheiden, ob und in welchem Umfang (Pos. 3, 4 ,5 ,6 zusammen für bis zu 50% des finanziellen Gesamtvolumens von Position 1 und 2) er diese optionale Position in Anspruch nimmt oder nicht. Angebotsrelevant ist der Preis für K zusätzliche GPU-Knoten, wobei K = 10% der GPU-Knoten ist, die in Position 1 angeboten wurden. Beachten Sie, dass dieser Preis rein fiktiv ist, da K auch nicht ganzzahlig sein kann. Dieser Preis für 10% zusätzliche GPU-Knoten wird zur Bestimmung des Gesamtpreises berücksichtigt, mit dem das wirtschaftlich günstigste Angebot ermittelt wird. (Mengeneinheit 1 Stück entspricht K=10%) Hinweis: Bei der zu erbringenden Leistung handelt es sich um eine optionale Position.	**Menge:** 1 Stück **Preiseinheit:** 1 Stück **Nettopreis in Euro** _____ **USt.:** 19 %, falls abweichend _____ % **Nachlass (%)** _____	_____

<table>
<tr><th>Nr.</th><th>Bezeichnung</th><th>Mengen- und Preisangaben</th><th>Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)</th></tr>
<tr><td>F 3.1</td><td colspan="2">zu 3: Fragebogen zu Option Mehrabnahme zum Zeitpunkt des Zuschlags (GPU-Knoten)
Da die Mehrkosten durch zusätzliche GPU-Knoten nicht allein mit der Knotenanzahl steigen, machen Sie bitte darüber hinaus folgende verbindliche Angaben zur Mehrkosten für benötigte zentrale Komponenten wie z.B. Chassis/Enclosures, CDUs, Switche und Racks. Diese Angaben dienen dazu, dass der AG wirtschaftlich und technisch sinnvolle Erweiterungsmöglichkeiten mit verbindlichen Kosten bestimmen kann. Dass heißt, es dürfen keine weiteren Kosten für Hardware, Software, Installation und Wartung durch die Mehrabnahme entstehen, als aus diesen Angaben hervorgeht.</td><td></td></tr>
<tr><td></td><td>Fragetitel</td><td>Antwort</td><td></td></tr>
<tr><td></td><td>1.1 Granularität zusätzliche GPU-Knoten
Bitte geben Sie an, in welcher Granularität N zusätzliche GPU-Knoten bestellt werden können - berücksichtigen Sie hier Einschränkungen, dass jeweils N Knoten eine Einheit/ein Chassis/ein Enclosure o.ä. bilden, so dass nur Vielfache von N sinnvoll und ökonomisch günstig angeboten werden können. Der AG ist an einer geringen Granularität, möglichst N=1, interessiert.</td><td></td><td></td></tr>
<tr><td></td><td>1.2 Erweiterungsknoten im angebotenen System ohne zusätzliche zentrale Komponenten
Welche Anzahl an GPU-Knoten kann dem unter Position 1 angebotenen System maximal hinzugefügt werden, bevor zusätzliche zentrale Komponenten installiert werden müssen, die nicht im genannten Preis der Knoten eingeschlossen sind?</td><td></td><td></td></tr>
<tr><td></td><td>1.3 Preise zusätzlicher zentralen Komponenten
Bitte nennen Sie die Einzelpreise zusätzlicher zentralen Komponenten die benötigt werden, um das System um weitere GPU-Knoten bis zum (in der Positionsbeschreibung genannten) Maximum zu erweitern, die nicht bereits im Preis der Knoten eingeschlossen sind.</td><td></td><td></td></tr>
<tr><td></td><td>1.4 Preis für einen zusätzlichen GPU-Knoten an.
Bitte geben Sie den Preis für einen zusätzlichen GPU-Knoten an.</td><td>Antwort - Betrag in Euro
Euro</td><td></td></tr>
</table>

<table>
<tr><th>Nr.</th><th>Bezeichnung</th><th>Mengen- und Preisangaben</th><th>Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)</th></tr>
<tr><td>4</td><td>Optionale Position - relevant für Angebotssumme
Wartung für das 4. und 5. Betriebsjahr für Mehrabnahme zum Zuschlag

Kosten der Verlängerung des Wartungs- und Instandhaltungszeitraums für das 4. und 5. Jahr für zusätzliche GPU-Knoten aus Position 3.

Angebotsrelevant ist der Preis für die Wartungsverlängerung für zwei Jahre von K zusätzlichen GPU-Knoten, wobei K = 10% der GPU-Knoten ist, die in Position 1 angeboten wurden. Beachten Sie, dass dieser Preis rein fiktiv ist, da K auch nicht ganzzahlig sein kann. Der Preis für die Wartungsverlängerung für zwei Jahre (4. und 5. Jahr zusammen) für 10% zusätzliche Knoten wird zur Bestimmung des Gesamtpreises berücksichtigt, mit dem das wirtschaftlich günstigste Angebot ermittelt wird. Diese Option kann durch den Auftraggeber nur innerhalb der ersten 3 Jahre nach Abnahme beauftragt werden. Bei der zu erbringenden Leistung handelt es sich um eine optionale Position.

Die Mindestanforderungen (=Ausschlusskriterien) an den Leistungsgegenstand ergeben sich aus Kriterienhauptgruppen (KHG) A bis H.

Hinweis: Bei der zu erbringenden Leistung handelt es sich um eine optionale Position.</td><td>Menge: 2 Jahr

Preiseinheit: 1 Jahr

Nettopreis in Euro

USt.: 19 %, falls abweichend _____ %

Nachlass (%)
_____</td><td></td></tr>
<tr><td>F 4.1</td><td colspan="2">zu 4: Info-Fragebogen</td><td></td></tr>
<tr><td></td><td>Fragetitel</td><td>Antwort</td><td></td></tr>
<tr><td></td><td>1.1 Preis pro ein Jahr Wartungsverlängerung eines GPU-Knotens

Bitte geben Sie den Preis pro ein Jahr Wartungsverlängerung eines GPU-Knotens an</td><td>Antwort - Betrag in Euro

Euro</td><td></td></tr>
</table>

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
5	Optionale Position - **relevant für Angebotssumme** Mehrabnahme zu einem späteren Zeitpunkt (GPU-Knoten) Weitere GPU-Knoten in identischer Konfiguration wie unter Position 1 entsprechend KHG A bis H inkl. Garantie mit Gewährleistung und Support für 3 Jahre (Laufzeit beginnt nach Abnahme der zusätzlichen GPU-Knoten). Der Auftraggeber kann bis zum 30.06.2026 entscheiden, ob und in welchem Umfang (Pos. 3, 4, 5, 6 zusammen für bis zu 50% des finanziellen Gesamtvolumens von Position 1 und 2) er diese optionale Position in Anspruch nimmt oder nicht. Angebotsrelevant ist der Preis für K zusätzliche GPU-Knoten, wobei K = 10% der GPU-Knoten ist, die in Position 1 angeboten wurden. Beachten Sie, dass dieser Preis rein fiktiv ist, da K auch nicht ganzzahlig sein kann. Dieser Preis für 10% zusätzliche GPU-Knoten wird zur Bestimmung des Gesamtpreises berücksichtigt, mit dem das wirtschaftlich günstigste Angebot ermittelt wird. (Mengeneinheit 1 Stück entspricht K=10%) Hinweis: Bei der zu erbringenden Leistung handelt es sich um eine optionale Position.	**Menge:** 1 Stück **Preiseinheit:** 1 Stück **Nettopreis in Euro** _____ **USt.:** 19 %, falls abweichend _____ % **Nachlass (%)** _____	_____

<table>
<tr><th>Nr.</th><th>Bezeichnung</th><th>Mengen- und Preisangaben</th><th>Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)</th></tr>
<tr><td>F 5.1</td><td colspan="2">zu 5: Fragebogen zu Option Mehrabnahme zu einem späteren Zeitpunkt (GPU-Knoten)
Da die Mehrkosten durch zusätzliche GPU-Knoten nicht allein mit der Knotenanzahl steigen, machen Sie bitte darüber hinaus folgende verbindliche Angaben zur Mehrkosten für benötigte zentrale Komponenten wie z.B. Chassis/Enclosures, CDUs, Switche und Racks. Diese Angaben dienen dazu, dass der AG wirtschaftlich und technisch sinnvolle Erweiterungsmöglichkeiten mit verbindlichen Kosten bestimmen kann. Dass heißt, es dürfen keine weiteren Kosten für Hardware, Software, Installation und Wartung durch die Mehrabnahme entstehen, als aus diesen Angaben hervorgeht.</td><td></td></tr>
<tr><td></td><td>Fragetitel</td><td>Antwort</td><td></td></tr>
<tr><td></td><td>1.1 Granularität zusätzliche GPU-Knoten
Bitte geben Sie an, in welcher Granularität N zusätzliche GPU-Knoten bestellt werden können - berücksichtigen Sie hier Einschränkungen, dass jeweils N Knoten eine Einheit/ein Chassis/ein Enclosure o.ä. bilden, so dass nur Vielfache von N sinnvoll und ökonomisch günstig angeboten werden können. Der AG ist an einer geringen Granularität, möglichst N=1, interessiert.</td><td></td><td></td></tr>
<tr><td></td><td>1.2 Erweiterungsknoten im angebotenen System ohne zusätzliche zentrale Komponenten
Welche Anzahl an GPU-Knoten kann dem unter Position 1 angebotenen System maximal hinzugefügt werden, bevor zusätzliche zentrale Komponenten installiert werden müssen, die nicht im genannten Preis der Knoten eingeschlossen sind?</td><td></td><td></td></tr>
<tr><td></td><td>1.3 Preise zusätzlicher zentralen Komponenten
Bitte nennen Sie die Einzelpreise zusätzlicher zentralen Komponenten die benötigt werden, um das System um weitere GPU-Knoten bis zum (in der Positionsbeschreibung genannten) Maximum zu erweitern, die nicht bereits im Preis der Knoten eingeschlossen sind.</td><td></td><td></td></tr>
<tr><td></td><td>1.4 Preis für einen zusätzlichen GPU-Knoten an.
Bitte geben Sie den Preis für einen zusätzlichen GPU-Knoten an.</td><td>Antwort - Betrag in Euro
Euro</td><td></td></tr>
</table>

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
6	Optionale Position - **relevant für Angebotssumme** Wartung für das 4. und 5. Betriebsjahr für Mehrabnahme zu späteren Zeitpunkt Kosten der Verlängerung des Wartungs- und Instandhaltungszeitraums für das 4. und 5. Jahr für zusätzliche GPU-Knoten aus Position 5. Angebotsrelevant ist der Preis für die Wartungsverlängerung für zwei Jahre von K zusätzlichen GPU-Knoten, wobei K = 10% der GPU-Knoten ist, die in Position 1 angeboten wurden. Beachten Sie, dass dieser Preis rein fiktiv ist, da K auch nicht ganzzahlig sein kann. Der Preis für die Wartungsverlängerung für zwei Jahre (4. und 5. Jahr zusammen) für 10% zusätzliche Knoten wird zur Bestimmung des Gesamtpreises berücksichtigt, mit dem das wirtschaftlich günstigste Angebot ermittelt wird. Diese Option kann durch den Auftraggeber nur innerhalb der ersten 3 Jahre nach Abnahme beauftragt werden. Bei der zu erbringenden Leistung handelt es sich um eine optionale Position. Die Mindestanforderungen (=Ausschlusskriterien) an den Leistungsgegenstand ergeben sich aus Kriterienhauptgruppen (KHG) A bis H. Wenn die GPU-Knoten aus dieser Position erst nach Abnahme des Systems aus Position 1 (und ggf. Position 3) abgenommen werden, behält sich der AG vor, die Wartungszeiträume für Position 6 kürzer als ein Jahr zu beauftragen, wobei der Preis entsprechend proportional zur Zeit reduziert wird (in Einheiten von Monaten). Damit soll erreicht werden, dass der Wartungszeitraum für alle GPU-Knoten dieser Beschaffung zur gleichen Zeit endet. Hinweis: Bei der zu erbringenden Leistung handelt es sich um eine optionale Position.	**Menge:** 2 Jahr **Preiseinheit:** 1 Jahr **Nettopreis in Euro** _____ **USt.:** 19 %, falls abweichend _____ % **Nachlass (%)** _____	_____

Nr.	Bezeichnung	Mengen- und Preisangaben	Gesamtbetrag netto inkl. Pos.-Nachlass (EUR)
F 6.1	zu 6: Info-Fragebogen		
	Fragetitel	**Antwort**	
	Frage 1.1 Bitte geben Sie die Wartungskosten pro Monat pro Knoten an.	**Antwort - Betrag in Euro** _____ Euro	
	1.2 Preis pro ein Jahr Wartungsverlängerung eines GPU-Knotens Bitte geben Sie den Preis pro ein Jahr Wartungsverlängerung eines GPU-Knotens an.	**Antwort - Betrag in Euro** _____ Euro	

Skonto

Ein angebotenes Skonto wird nur berücksichtigt, wenn als Zahlungsziel mindestens 14 Tage angegeben werden!

1. Gewährung von _____ % Skonto bei Zahlung innerhalb von _____ Tagen

2. Gewährung von _____ % Skonto bei Zahlung innerhalb von _____ Tagen

Wertungsschema

UfAB-2018-Wertungsschema

Die Wertung erfolgt nach der einfachen Richtwertmethode nach UfAB 2018 (abrufbar unter http://www.cio.bund.de). Für die Bestimmung des wirtschaftlichsten Angebotes wird das Leistungs-Preis-Verhältnis herangezogen. Es wird jeweils der Quotient aus Leistungspunkten und Preis berechnet. Die so ermittelte Kennzahl wird mit dem Skalierungsfaktor 100000 multipliziert. Das Angebot mit dem höchsten Ergebnis wird als das wirtschaftlichste angesehen; bei mehreren Angeboten mit absolut gleichen Ergebnissen erhält das preisgünstigste den Zuschlag.

Summe der Gewichtungspunkte (GP): 10000 Gewichtungspunkte (GP)

Hinweis zur Darstellung des Erfüllungsgrades der Ausschluss- und Bewertungskriterien:

Bitte beantworten Sie für die jeweiligen Ausschluss- und Bewertungskriterien, in wieweit die von Ihnen angebotene Leistung die nachfolgenden Ausschluss- und Bewertungskriterien erfüllt. Bitte berücksichtigen Sie bei Ihrer Angebotslegung, dass bei der Beurteilung der Bewertungskriterien nur die Informationen berücksichtigt werden können, die Sie uns mit Ihrem Angebot bereitstellen.

Bei der Wertung der Angebote werden pro Kriterium 0 bis 10 (Bewertungs-)Punkte je nach Zielerfüllungsgrad vergeben:

8 bis 10 Punkte: hoher Zielerfüllungsgrad
4 bis 7 Punkte: durchschnittlicher Zielerfüllungsgrad
0 bis 3 Punkte: geringer Zielerfüllungsgrad

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
KHG A	**GPU-Compute-Knoten**		**5.800,00 GP**
A 1	**Mindestanforderungen** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Die formulierten Anforderungen der Positionen sowie die Vortexte sind im Sinne von Ausschlusskriterien (k.o.-Kriterien) zu verstehen. Angebote, die die Anforderungen (Ausschlusskriterien) nicht im vollen Umfang erfüllen, können für den Zuschlag nicht berücksichtigt und müssen ausgeschlossen werden.		
A 2	**Budgetgrenze** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Angebote, bei denen Position 1 das Budget in Höhe von 8.200.000,00 Euro (brutto, inkl. MwSt.) überschreiten, können für den Zuschlag nicht berücksichtigt werden.		
A 3	**Rechenleistung als wesentlicher Zweck** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Wesentlicher Zweck der Beschaffung ist die Bereitstellung von Rechenleistung auf GPUs für KI-Anwendungen für die Nutzer:innen des NHR-Zentrums der TU Dresden für 5 Jahre Laufzeit. Die Auswahl der angebotenen Hardware zielt auf die Maximierung der bereitgestellten Rechenleistung. Im Falle von langwierigen Verzögerungen oder Ausfällen bei der Bereitstellung der Hardware ist der Bereitstellung geeigneter Rechenleistung der Vorzug zu geben gegenüber der Einhaltung von Produkt- oder Hersteller-Angaben. Verzögerungen des Beginns des Produktivbetriebs des Systems, d.h. Start nach dem 1.10.2026, sind durch zusätzliche, identische GPU-Knoten zu kompensieren, so dass über die Betriebszeit von 5 Jahren, d.h. bis 30.09.2031, mindestens die gleiche Anzahl an GPU-Rechenstunden in Summe über diesen Zeitraum zur Verfügung steht. Eine möglichst frühzeitige Inbetriebnahme der zusätzlichen GPU-Knoten zur Kompensation wird bevorzugt, um die Anzahl der zusätzlichen GPU-Knoten zu minimieren.		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Kann ein Bieter nach sorgfältiger Betrachtung der zeitlichen Verfügbarkeit von Komponenten eine Verzögerung im Sinne des obigen Absatzes nicht vermeiden, sollte der Bieter die Kompensation bereits einplanen und offen im Angebot kommunizieren. Die zusätzlichen GPU-Knoten, die zur Kompensation geliefert werden, sind selbstverständlich nicht in die Leistungsbewertung einzubeziehen. Verzögerungen die mit dieser Kompensationsregelung kompensiert werden, unterliegen nicht der Vertragsstrafe nach EVB-IT-System-AGB. Kommt jedoch zum somit geplanten und mit dem AG abgestimmten, verzögerten Zeitplan weiterer Verzug hinzu, kann die Vertragsstrafe auf diese zusätzliche Verzögerung Anwendung finden. Der AG weist darauf hin, dass eine Verzögerung der Abnahme über das Jahr 2026 hinaus mit einer Bankbürgschaft (auf Kosten des Bieters) für die Abschlussrechnung verbunden ist.		
A 4	**GPU-Knoten** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Es ist ein Cluster aus GPU-Knoten anzubieten, wobei all diese Knoten identisch ausgestattet sein müssen. Die Knoten müssen mit dedizierten Host-CPUs ausgestattet sein. Sofern durch die Hardware ermöglicht, müssen die einzelnen GPUs durch Software logisch in mehrere Instanzen mit definierbaren Ressourcen je Instanz aufteilbar sein, damit unterschiedliche Anwendungen unterschiedlicher User gleichzeitig auf einer GPU verarbeitet werden können. Die Anzahl logischer Instanzen muss vom AG konfigurierbar sein. Wie viele GPUs mit wie vielen Instanzen initial eingerichtet werden, wird vor Inbetriebnahme des Systems mit dem AG abgestimmt. Die nötige Software, gegebenenfalls nötige Lizenzen, eine initiale Konfiguration sowie die Integration in Slurm sind durch die Bieter für die Betriebsjahre des Systems zu stellen. Die CPUs müssen der neuesten CPU-Generation des gewählten Herstellers		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	angehören, die zum Zeitpunkt der Angebotserstellung allgemein verfügbar ist. GPUs müssen ebenfalls der neuesten GPU-Generation des Herstellers angehören, die zum Zeitpunkt der Angebotserstellung allgemein verfügbar ist. Es dürfen auch CPU- oder GPU-Generationen angeboten werden, die erst zwischen Angebotserstellung und der vorgesehenen Installation verfügbar werden. Es sind Dual-Sockel-Systeme mit 2 Host-CPUs anzubieten. Die Knoten müssen mindestens 12 physikalische CPU-Kerne je GPU für Batchjobs zur Verfügung stellen. Dass heißt, CPU-Kerne die für Betriebssystem, Dateisystem oder andere Zwecke fest zugewiesen werden und nicht für Compute-Jobs zur Verfügung stehen, können für diese Anforderung nicht einberechnet werden. Jeder Compute-Knoten muss mindestens 256 GB Host-RAM je GPU beinhalten. Der Host-RAM muss dem höchsten Standard, der von der CPU unterstützt wird, entsprechen, mindestens jedoch DDR5-5600. In jedem Compute-Knoten ist eine Data-Center-NVMe-Disk (ein Device, kein RAID) einzurichten, die für das Betriebssystem und temporären Speicher genutzt werden soll. Dabei sollen pro verbauter GPU jeweils mindestens 1800 GB von diesem lokalen temporären Speicher verfügbar sein. Jeder GPU-Knoten ist mit einer Bandbreite von mindestens 200 Gbit/s pro GPU in einem HPC-Interconnect anzubinden. Die Compute-Knoten müssen 1 Höheneinheit pro GPU oder weniger einnehmen, wobei ein Verschnitt nicht mitberechnet wird, wenn das letzte Chassis nicht voll bestückt ist o.ä. Alle Compute-Knoten oder Chassis für Compute-Knoten benötigen ausdrücklich keine redundanten Netzteile.		
B 5	**CPU-Architektur** Werden für die GPU-Knoten Host-CPUs angeboten, die den x86-64 Instruktions-Set beherrschen, werden für dieses Kriterium		500 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	10 Bewertungspunkte vergeben, andernfalls 0. Hinweis: die CPUs der Service-Knoten aus KHG B müssen der gleichen Mikroarchitektur angehören, wie die CPUs der GPU-Knoten.		
B 6	**Theoretische Rechenleistung der GPUs** Die theoretische Rechenleistung der GPUs laut Spezifikation ist ein wichtiger Indikator für die gebotene Rechenleistung, auch wenn sie als theoretische obere Schranke nicht für reale Anwendungen zu erwarten ist. Zur Beurteilung der theoretischen Rechenleistung wird die Rate folgender Arten von Operationen betrachtet: FP8-Tensor, BF16-Tensor, TF32-Tensor, FP64-Vector. Für die genannten Operationen wird pro GPU ein gewichtetes Mittel gebildet wie folgt: RateMittel = 1/8 * RateFP8-Tensor + 1/4 * RateBF16-Tensor + 1/2 * RateTF32-Tensor + RateFP64-Vector Bei der Festlegung der Rechenleistung für Tensor-Operationen ist von dichtbesetzten Matrizen auszugehen. Wenn keine dedizierten Tensor-Operationen für die Zahlentypen vorhanden sind, dürfen Raten für Tensor-Operationen der nächstgrößeren Zahlentypen verwendet werden oder allgemeine (nicht-Tensor) Raten für die gleichen Zahlentypen. Wenn keine FP64-Vector-Operationen zur Verfügung stehen, muss RateFP64-Vector = 0 in obiger Gleichung eingesetzt werden. Diese Rate wird mit der Anzahl N von insgesamt angebotenen GPUs aller GPU-Knoten multipliziert und ergibt: RateGesamt = N * RateMittel Anhand RateGesamt sind maximal 10 Bewertungspunkte zu erreichen: 1 Punkt für RateGesamt >= 158 PFlop/s 2 Punkte für RateGesamt >= 181 PFlop/s 3 Punkte für RateGesamt >= 204 PFlop/s 4 Punkte für RateGesamt >= 227 PFlop/s 5 Punkte für RateGesamt >= 250 PFlop/s		1.500 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	6 Punkte für RateGesamt >= 273 PFlop/s 7 Punkte für RateGesamt >= 296 PFlop/s 8 Punkte für RateGesamt >= 319 PFlop/s 9 Punkte für RateGesamt >= 342 PFlop/s 10 Punkte für RateGesamt >= 365 PFlop/s		
B 7	**Host- und GPU-RAM** Die Speicherausstattung von GPUs ist ein entscheidender Faktor für den effizienten Einsatz großer KI-Modelle. Höhere Speicherkapazitäten erlauben eine Reduktion der Anzahl benötigter GPUs und vereinfachen die Modellverarbeitung. Bewertet werden sowohl der verfügbare GPU-Speicher als auch der Hauptspeicher (RAM) des Hostsystems. Mehr Host-RAM ist von Vorteil und der AG bevorzugt Konfigurationen, bei denen der Host-RAM mindesten der Summe aller GPU-Speicherkapazitäten in den Knoten entspricht (der AG behält sich vor dies nach den Verhandlungen zu fordern bzw. zu konkretisieren). Für eine möglichst hohe Speicherbandbreite des Host-RAM müssen alle Speicherkanäle voll bestückt sein mit DIMMs gleicher Kapazität. Eine Doppelbestückung der Speicherkanäle ist zulässig, wenngleich der Auftraggeber einer Einfachbestückung bevorzugt. Insgesamt sind 10 Bewertungspunkte zu erreichen. Anhand der Größe des GPU-Speichers je GPU sind maximal 5 Bewertungspunkte zu erreichen: 1 Punkt für GPU-Speicher >= 160 GB 2 Punkte für GPU-Speicher >= 192 GB 3 Punkte für GPU-Speicher >= 224 GB 4 Punkte für GPU-Speicher >= 256 GB 5 Punkte für GPU-Speicher >= 288 GB Anhand der Größe des Host-RAMs je GPU sind maximal weitere 5 Bewertungspunkte zu erreichen: 1 Punkt für Host-RAM je GPU >= 256 GB 2 Punkte für Host-RAM je GPU >= 288 GB 3 Punkte für Host-RAM je GPU >= 320 GB 4 Punkte für Host-RAM je GPU >= 352 GB 5 Punkte für Host-RAM je GPU >= 384 GB		200 GP
B 8	**Benchmark-Leistung mit HPL-MxP** Für die Bewertung der Rechenleistung ist der HPL-MxP-Benchmark (High		1.100 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Performance Linpack Mixed Precision) über alle angebotenen GPU-Knoten auszuführen. Er ist in einer Weise auszuführen, dass alle GPU-Knoten beteiligt werden und alle GPUs genutzt werden. Der erreichte Score ist anzugeben. Die verwendete Benchmark-Konfiguration sowie alle Angaben für eine Einreichung des Ergebnisses auf https://hpl-mxp.org/ entsprechend der dort veröffentlichten Regeln ist anzugeben. Basierend auf den Angaben soll der AG in der Lage sein, den Benchmarklauf und die erzielte Leistung zu reproduzieren. Anhand der Leistung mit HPL-MxP können maximal 10 Leistungspunkte erreicht werden: 1 Punkt für HPL-MxP-Leistung >= 80 PFlop/s 2 Punkte für HPL-MxP-Leistung >= 100 PFlop/s 3 Punkte für HPL-MxP-Leistung >= 120 PFlop/s 4 Punkte für HPL-MxP-Leistung >= 140 PFlop/s 5 Punkte für HPL-MxP-Leistung >= 160 PFlop/s 6 Punkte für HPL-MxP-Leistung >= 180 PFlop/s 7 Punkte für HPL-MxP-Leistung >= 200 PFlop/s 8 Punkte für HPL-MxP-Leistung >= 220 PFlop/s 9 Punkte für HPL-MxP-Leistung >= 240 PFlop/s 10 Punkte für HPL-MxP-Leistung >= 260 PFlop/s Falls die angebotenen GPUs einen gültigen Lauf des HPL-Benchmarks (https://top500.org/project/linpack/) zulassen, kann der AG verlangen, dass der AN diesen Benchmark zur Abnahme ebenfalls durchführt und die Konfiguration angibt. Wertungsrelevant für dieses Kriterium ist jedoch nur der HPL-MxP-Benchmark.		
B 9	**Benchmark-Leistung mit DeepSpeed-DNN-Benchmark** Für die Bewertung der Rechenleistung für KI-Anwendungen ist das Trainieren eines BERT-Modells durchzuführen. Der Benchmark ist unter [0] hinterlegt. Als		1.500 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Datensatz ist ein Wiki-Corpus mit 56GB unter [2] hinterlegt. Der Benchmark ist auf insgesamt 16 GPUs auszuführen, wobei alle GPUs der beteiligten Knoten genutzt werden sollen. Falls die Anzahl der GPUs beteiligter Knoten kein Teiler von 16 ist, müssen mehr (aber nicht weniger) GPUs genutzt werden, um die beteiligten Knoten voll auszulasten. Änderungen des Quellcodes im Sinne der Performance-Optimierungen und der Hyperparameter, mit Ausnahme der Micro-Batchsize, sind unzulässig. Quellcode-Änderungen, welche der Funktionsfähigkeit der Anwendung dienen (z.B. Bugs) sind bekannt zu machen. Es sind mindestens Pytorch in der Version 2.6 und das DeepSpeed-Framework [3] in der Version 0.16.5 zu verwenden. Hardwarespezifische Pytorch-Versionen müssen frei verfügbar sein. Die Veränderung der Micro-Batchsize erfolgt durch Veränderung des Wertes "train_micro_batch_size_per_gpu" in der Datei "deepspeed_bsz64k_lamb_config_seq128_tud.json". Eine Installations-Anleitung wird in der Datei "INSTALL.md" [1] bereitgestellt. Der Datensatz soll auf dem gelieferten Speichersystem (siehe KHG C) platziert und nicht manuell auf die Knoten kopiert werden. Automatische Cache-Mechanismen, die im Produktionsbetrieb vorgesehen sind, sind hingegen gestattet. Die verwendete Benchmark-Konfiguration mit den verwendeten Softwarebibliotheken ist anzugeben, so dass der AG in der Lage ist, den Benchmarklauf und die erzielte Leistung zu reproduzieren. Zur Bewertung der Leistung im DNN-Benchmark wird die Laufzeit T_DNN in Sekunden ermittelt, die benötigt wird, um einen Train-Loss von 6.0 für mindestens 327680 Samples bzw. 5 Trainings-Iterationen zu unterschreiten. Zur Ermittlung der Laufzeiten und Loss-Werte ist das Skript "export_from_tensorboard.py" zu verwenden. Dort stehen sowohl die Spalten Loss und Runtime als auch die konkrete Zeitangabe am Ende der Ausgabe bereit.		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Die Zeitmessung beginnt nach Berechnung der ersten Trainings-Iteration. Für die Wertung relevant ist der Durchsatz auf allen GPU-Knoten des gleichen Typs (im Sinne von Anzahl Benchmark-Instanzen pro Stunde) der wie folgt berechnet wird: Durchsatz = (3600s / T_DNN) * (N_Gesamt / N_DNN) Wobei N_Gesamt die Anzahl der insgesamt angebotenen GPUs ist und N_DNN die Anzahl der für diesen Benchmark genutzten GPUs ist. Der Durchsatz kann ein nicht-ganzzahliger Wert sein. Auf dieser Basis sind max. 10 Punkte erreichbar: 1 Punkt für Durchsatz >= 7,00 Instanzen / h 2 Punkte für Durchsatz >= 8,00 Instanzen / h 3 Punkte für Durchsatz >= 9,00 Instanzen / h 4 Punkte für Durchsatz >= 10,00 Instanzen / h 5 Punkte für Durchsatz >= 11,00 Instanzen / h 6 Punkte für Durchsatz >= 12,00 Instanzen / h 7 Punkte für Durchsatz >= 13,00 Instanzen / h 8 Punkte für Durchsatz >= 14,00 Instanzen / h 9 Punkte für Durchsatz >= 15,00 Instanzen / h 10 Punkte für Durchsatz >= 16,00 Instanzen / h [0] https://github.com/tud-zih-ki/DeepSpeedExamples/tree/gpu_bench_deneb/training/bing_bert [1] https://github.com/tud-zih-ki/DeepSpeedExamples/blob/gpu_bench_deneb/training/bing_bert/INSTALL.md [2] https://cloudstore.zih.tu-dresden.de/index.php/s/RyeDoaogesGJFCR		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	[3] https://github.com/deepspeedai/DeepSpeed		
B 10	**Benchmark-Leistung mit vLLM-Inferenz** Zur Leistungsbewertung der angebotenen Systeme bzgl. der effizienten Ausführung großer Sprachmodelle (LLMs) im Inferenzbetrieb wird ein standardisierter Inference-Benchmark unter Verwendung des Open-Source-Frameworks vLLM (https://github.com/vllm-project/vllm) durchgeführt. Das Ziel ist es, ein System bereitzustellen, das hohe Inferenzgeschwindigkeiten bei gleichzeitig niedriger Latenz und optimaler GPU-Auslastung ermöglicht. Für die Durchführung des Inferenz-Benchmarks mit vLLM gelten folgende verbindliche Vorgaben bzgl. der Softwareumgebung: • vLLM-Version: Es ist ausschließlich die Version vLLM v0.9.1 zu verwenden. Code-Modifikationen sind nicht zulässig. • Containerisierung: Die Tests sind innerhalb eines Container-Setups durchzuführen. Sowohl AMD ROCm-basierte als auch NVIDIA CUDA-basierte Container-Umgebungen sind zulässig. Die Wahl der konkreten Container-Images (z. B. PyTorch-Version, CUDA/ROCm-Version) ist freigestellt, muss jedoch im Angebot eindeutig angegeben und dokumentiert werden. • vLLM V1-Modus: Die Benchmarks sind mit aktivierter V1-Architektur durchzuführen. Hierzu ist beim Start die Umgebungsvariable VLLM_USE_V1=1 zu setzen. Abweichungen von diesen Vorgaben führen zur Ungültigkeit der Benchmark-Ergebnisse. Es ist verpflichtend das Modell meta-llama/Llama-4-Scout-17B-16E-Instruct in unveränderter Form (kein Finetuning, keine Quantisierung oder Modifikation) zu verwenden. Das Modell ist direkt aus dem Hugging Face Model Hub zu laden (https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct) oder in der identischen, originalen Version lokal		1.000 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	bereitzustellen. Die Durchführung der Benchmarks erfolgt mittels des Online-Inferenzservers von vLLM sowie der zugehörigen Benchmarking-Suite. Dabei sind folgende verbindliche Konfigurationen und Abläufe einzuhalten. • Der vLLM-Inferenzserver ist mit folgender Konfiguration zu starten, alle nicht explizit genannten Parameter sind im Default-Zustand zu belassen und dürfen nicht modifiziert werden: vllm serve meta-llama/Llama-4-Scout-17B-16E-Instruct --port 8000 --max-model-len 1M --kv-cache-dtype fp8 --tensor-parallel-size 4 --disable-log-requests • Als Lastgenerator ist die offizielle vLLM-Benchmarking-Suite zu verwenden, mit folgender Konfiguration: vllm bench serve --base-url http://127.0.0.1:8000 --model meta-llama/Llama-4-Scout-17B-16E-Instruct --dataset-name random --random-input-len $ISL --random-output-len $OSL --max-concurrency $CONCURRENCY --num-prompts 12800 --ignore-eos --percentile_metrics ttft,tpot,itl,e2el Es ist ein festes Szenario mit einer Eingabesequenzlänge (ISL) von 1000 Tokens und einer Ausgabelänge (OSL) von 1000 Tokens durchzuführen (Verhältnis 1:1). Der einzige zu variierende Parameter ist $CONCURRENCY, welcher die gleichzeitige Anzahl aktiver Anfragen steuert. Die Testdurchführung erfolgt mit frei wählbaren $CONCURRENCY-Werten, um systemseitig die bestmögliche Performance zu ermitteln. Dabei darf Median TTFT (Time to First Token) nicht über 1500ms liegen. Im Rahmen des Abnahme sind die Konfiguration, Durchführung und Ergebnisse des Benchmark vollständig zu dokumentieren. Dabei sind folgende Leistungskennzahlen (Metriken) für die Benchmarkkonfiguration verpflichtend anzugeben:		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	• Throughput-Metriken: Output Token Throughput (OTT) in Tokens pro Sekunde, Total Token Throughput (TTT) in Tokens pro Sekunde und Request Throughput (req/s). • Latenz-Metriken (jeweils in Millisekunden, in den Ausprägungen mean, median, P99): Time to First Token (TTFT), Time per Output Token (TPOT), Inter-Token Latency (ITL) und End-to-End Latency (E2EL) Zur Abnahme des GPU-Clusters ist ein skalierter Inferenz-Benchmark mit vLLM gemäß den folgenden Vorgaben durchzuführen: Auf allen GPU-Knoten des Clusters sind zeitgleich Benchmark-Instanzen zu starten. Jede Instanz nutzt exakt 4 GPUs, so dass alle im Cluster vorhandenen GPUs vollständig ausgelastet sind (d. h. N/4 Instanzen insgesamt, wobei N die Anzahl der installierten GPUs im Cluster ist). Jede Instanz wird separat ausgeführt, mit identischer Modellkonfiguration und Benchmark-Setup gemäß den oben beschriebenen Anforderungen. Für die Abnahme gilt die Instanz mit der niedrigsten erreichten Total Token Throughput (minTTT in Requests/s) über den gesamten Durchlauf als maßgeblich, wobei Median TTFT (Time to First Token) bei keiner Instanz über 1500 ms liegen darf. Der endgültige Abnahmewert wird berechnet als: Effektive Requests/s = minTTT * (N / 4). Anhand der zugesagten Effektive Requests/s sind für diese Kriterium bis zu 10 Wertungspunkte wie folgt erreichbar: 1 Punkt für >= 200 Requests/s 2 Punkte für >= 230 Requests/s 3 Punkte für >= 260 Requests/s 4 Punkte für >= 290 Requests/s 5 Punkte für >= 320 Requests/s 6 Punkte für >= 350 Requests/s 7 Punkte für >= 380 Requests/s 8 Punkte für >= 410 Requests/s 9 Punkte für >= 440 Requests/s 10 Punkte für >= 470 Requests/s		
A 11	**Benchmark-Leistung mit**		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Kommunikations-Benchmark (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Für die Bewertung der maximalen Transferraten zwischen den GPUs ist ein Allreduce-Benchmark auszuführen. Im Fall von GPUs des Herstellers Nvidia ist der Benchmark nccl-tests [0] und im Fall von GPUs des Herstellers AMD ist der Benchmark rccl-tests [1] auszuführen. Sollten die angebotenen GPUs weder dem Hersteller Nvidia noch AMD zugeordnet werden können, sind die Bandbreiten vom Hersteller mittels eines selbst gewählten Benchmarks nachzuweisen. Der Benchmark muss dabei analog zu dem Benchmark nccl-tests die Allreduce-Operation in gleicher Art und Weise ausführen. Die Allreduce-Operation ist dabei mit geeigneten Bibliotheken auszuführen, wie sie auch in Pytorch unter realen Anwendungen Verwendung finden. Änderungen des Quellcodes (nccl-tests, rccl-tests) im Sinne der Performance-Optimierungen und der Hyperparameter, mit Ausnahme der Nachrichtengröße, sind unzulässig. Quellcode-Änderungen, die der Funktionsfähigkeit der Anwendung dienen (z. B. Bugs, zusätzliche Kommunikations-Bibliotheken), sind bekannt zu machen. Maßgeblich zur Erfüllung dieses Kriteriums sind die minimalen Bandbreiten in der Konfiguration mit allen GPUs eines Compute-Knotens (knoteninterne Kommunikation) und die minimalen Bandbreiten mit allen GPUs von 4 Compute-Knoten (inter-Knoten-Kommunikation). Die Leistungswerte müssen repräsentativ sein, d. h. sie dürfen auf keinem Knoten bzw. keiner Teilmenge von Knoten mehr als 5% nach unten abweichen. Es müssen mindestens 150 GB/s für die knoteninterne Kommunikation und 75 % des theoretischen Maximums der angebotenen HPC-Interconnect-Bandbreite (unidirektional) eines Knotens für die inter-Knoten-Kommunikation erreicht werden. Maßgeblich für dieses Kriterium ist der Wert für "Bus Bandwidth" der Benchmarks (Spalte "busbw" in nccl-tests bzw. rccl-tests).		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Die verwendete Benchmark-Konfiguration mit den verwendeten Softwarebibliotheken ist anzugeben, so dass der AG in der Lage ist, den Benchmarklauf und die erzielte Leistung zu reproduzieren.		
A 12	**Unterstützung von APIs und Frameworks für HPC und KI** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Für alle angebotenen GPU-Knoten muss die Programmierung der GPUs in CUDA oder einer in Syntax und Semantik daran angelehnten API möglich sein. Sofern CUDA nicht nativ genutzt werden kann, müssen Werkzeuge bereitgestellt sein, welche die Konvertierung von in CUDA geschriebenen Programmen in die alternative, daran angelehnte API ermöglichen. Die API muss mit dem Message Passing Interface (MPI) genutzt werden können, so dass mehrere GPUs, auch über mehrere Compute-Knoten hinweg, genutzt werden können. Zudem wird die Unterstützung von OpenMP-Offloading mindestens nach Standard 5.0 vorausgesetzt, wofür Compiler zur Verfügung gestellt werden müssen. Die API sowie OpenMP-Offloading müssen mit C/C++ sowie Fortran nutzbar sein. Werkzeuge für die Analyse der Performance (Profiling, Tracing) von GPU-Anwendungen müssen für Nutzende kostenfrei verfügbar sein. Die Analyse der durch das Profiling oder Tracing erzeugten Daten muss mit den zur Verfügung gestellten Werkzeugen auch auf anderen als diesem System möglich sein. Die dafür genutzten Werkzeuge müssen auch dann nutzbar sein, wenn keine GPU des Herstellers vorhanden ist, d.h. mindestens auch auf anderen Knoten in der HPC-Umgebung. Die GPU-Programmierschnittstelle (API) muss, sofern der Auftraggeber dies im Rahmen des Leistungsabrufs verlangt, während des gesamten Betriebszeitraums vollständig mit den folgenden Frameworks kompatibel sein und deren Nutzung ermöglichen: PyTorch (einschließlich Torchvision und Torchaudio) [0], JAX [1], TensorFlow [2] sowie vLLM [3].	☐ **Ja** ☐ **Nein**	

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Sofern herstellerspezifische Varianten dieser Frameworks erforderlich sind, müssen diese spätestens vier Monate nach Veröffentlichung der jeweiligen frei verfügbaren Community-Version auf dem angebotenen System verfügbar, installierbar und lauffähig sein. Die APIs sowie sämtliche Tools müssen öffentlich dokumentiert sein. [0] https://pytorch.org [1] https://docs.jax.dev/en/latest/ [2] https://www.tensorflow.org/ [3] https://blog.vllm.ai/2023/06/20/vllm.html		
KHG B	**Service-Knoten**		**200,00 GP**
A 13	**Service-Knoten** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Es sind 8 identisch ausgestattete Service-Knoten mit adäquater Rechenleistung für Login und datenintensives Prä- bzw. Postprocessing (ohne GPUs) anzubieten. Die CPUs der Service-Knoten sollen identische Mikroarchitektur haben wie die Host-CPUs der GPU-Knoten (KHG A). Die Service-Knoten sollen über mindestens 100 CPU-Cores, 2 TB Hauptspeicher und 15 TB lokalen NVMe-Speicher verfügen. Jeder Knoten ist mit einem Port mit einer Bandbreite von mindestens 200 Gbit/s im selben HPC-Interconnect integriert wie die GPU-Knoten. 4 Knoten davon verfügen darüber hinaus über einen weiteren Ethernet-Port mit mindestens 200 Gbit/s, der aber nicht im Rahmen dieses Angebots angeschlossen wird.		
B 14	**Rack-Integration** Werden diese Service-Knoten in den Racks und derselben Infrastruktur (Warmwasserkühlung, siehe A29) wie die GPU-Knoten betrieben, wird das positiv bewertet (10 Punkte). Alternativ ist eine Installation im Raum E15 (vgl. Anlage 03 - "Stellplan", Stellfläche 8) möglich (0 Punkte).		200 GP
KHG C	**Storage**		**200,00 GP**

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
A 15	**Speichersystem** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Es ist ein nichtflüchtiges, flashbasiertes Speichersystem mit einer nutzbaren Nettokapazität von mindestens 2 PB anzubieten. Dieses Speichersystem soll als schneller Scratch-Speicher dediziert für das GPU-Cluster (siehe KHG D - Netzwerke) dienen. Als Installationsort für das Speichersystem ist der Raum E15 (vgl. Anlage 03 "Stellplan", Stellfläche 7) vorgesehen. Der Speicher soll von den Nutzenden vor Beginn der Experimente mit entsprechenden Eingaben aus bereits vorhandenen Speichersystemen befüllt werden. Anschließend lesen und schreiben die Nutzenden während der Experimente auf dem GPU-Cluster ihre Daten aus und in das Scratch-Speichersystem. Ein Experiment kann dabei auch aus mehreren einzelnen Jobs bestehen und erfordern, dass die Daten mehrere Tage bis Wochen gespeichert werden. Nach Beendigung des Experiments sollen die Daten wieder zurück in ein anderes Speichersystem überführt werden. Das Kopieren der Daten zwischen den Speichersystemen ist als Anwendungsfall zu berücksichtigen.		
A 16	**Dateisystem für Speicher** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Zum Betreiben des Speichersystems ist ein paralleles Dateisystem mit auszuliefern, welches für den parallelen Zugriff von mehreren Knoten ausgelegt ist. Das angebotene Dateisystem muss zwingend GPU-Direct-Storage (GDS) unterstützen, unabhängig vom Hersteller und Modell der gelieferten GPUs. Das heißt, dass Dateisystem muss ermöglichen, dass GPUs für Datenzugriffe direkt über Remote-Direct-Memory-Access (RDMA) auf den Speicher des Speichersystems zugreifen können, ohne das Daten zusätzlich durch den Speicher der CPU kopiert werden müssen. Das Dateisystem muss mindestens eine POSIX-I/O Schnittstelle anbieten, die vollständig unterstützt wird. Den Nutzenden soll ein hierarchischer Namensraum in Form eines Verzeichnisbaums zum		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Arbeiten mit ihren Dateien angeboten werden. Das Dateisystem sollte auf den Betrieb von Flash-basierten Speichersystemen optimiert sein. Außerdem muss das Storage-System Statistiken pro Slurm-Job liefern können. Die Mindestanforderung bzgl. der Metriken sind aggregierte IOPs sowie aggregierte Lese- und Schreib-Bandbreite. Weitere Statistiken, die Einsichten in problematische I/O-Muster und Ungleichgewichte im Storage-System erlauben, wie z.B. Anzahl der Metadaten-Operationen, Lese- und Schreibraten per OST, durchschnittliche I/O-Größe per OST, etc. sind wünschenswert, aber optional. Das Speichersystem muss Mechanismen zur Umsetzung von Quality-of-Service-Maßnahmen (QoS) bereitstellen, um eine faire Ressourcennutzung und die Vermeidung von Performance-Engpässen durch einzelne Nutzer oder Clients zu gewährleisten. Systeme, die eine feingranulare Kontrolle pro Nutzer, Client ermöglichen - z. B. durch Begrenzung von Bandbreite (MB/s) und/oder IOPS - werden bevorzugt. Die QoS-Maßnahmen müssen dynamisch konfigurierbar und zur Laufzeit ohne Neustart des Systems anpassbar sein. Die Konfiguration von QoS-Regeln soll über standardisierte Schnittstellen oder APIs möglich sein. Das Angebot enthält technische Erläuterungen zur Umsetzung dieser Anforderungen.		
B 17	**Features des Dateisystems** Zusätzlich zu den in A16 (Dateisystem für Speicher) definierten Anforderungen wird das angebotene Dateisystem hinsichtlich folgender optionaler Funktionalitäten bewertet. Für alle Funktionalitäten zusammen werden maximal 10 Wertungspunkte vergeben: Die native Unterstützung des S3-Protokolls (Simple Storage Service) durch das Dateisystem wird wie folgt bewertet (max. 5 Punkte): • 5 Punkte: Vollständige native S3-Unterstützung ohne zusätzliche Gateway-Komponenten oder Mounts, mit		200 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	dokumentierter Performance, Authentifizierung und Bucket-Management. • 3 Punkte: S3-Zugriff über integriertes Gateway oder Zusatzmodule des Herstellers, mit funktionaler Integration. • 1 Punkt: Grundlegende S3-Konnektivität mit Einschränkungen oder durch externe Lösungen. Die Fähigkeit, Software-Updates (z. B. sicherheitsrelevante Patches oder Funktionsaktualisierungen) im laufenden Nutzerbetrieb für das Dateisystem durchzuführen, wird wie folgt bewertet (max. 5 Punkte): • 5 Punkte: Nachweislich unterstützte unterbrechungsfreie Updates oder vergleichbare Verfahren ohne Betriebsunterbrechung für Client- und Server-Komponenten, inkl. Herstellerdokumentation und Referenzinstallation. • 3 Punkte: Updates für Client- und Server-Komponenten sind im laufenden Nutzerbetrieb möglich, aber mit kurzzeitigen Einschränkungen (z. B. temporärer Performanceeinbruch oder Dienstneustarts). • 1 Punkt: Alle Client-Komponenten (z. B. Kernel-Module, Treiber oder User-Space-Tools) können ohne Reboot aktualisiert werden. Die Server-Updates erfordern geplante Downtimes, aber mit automatisierter Unterstützung. Die Gesamtbewertung erfolgt auf Basis der vom Anbieter eingereichten Unterlagen (z. B. technische Dokumentation, Herstellerangaben, Erfahrungsberichte) und kann durch eine Teststellung validiert werden. Alle Funktionalitäten, die das angebotene Dateisystem erfüllt, sind in der Abnahme nachzuweisen.		
A 18	**Abnahmetests für das Speichersystem** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Die Bandbreite des Speicherssystem muss für den lesenden parallelen I/O pro GPU-Knoten bei einzelner Messung mit nur einem Knoten mindestens eine Bandbreite von G * 50 GBit/s erreichen, wobei G die Anzahl der GPUs pro Knoten ist. Für alle GPU-Knoten gleichzeitig müssen mindestens 90% von K * 50 GBit/s erreicht		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	werden, wobei K die Anzahl aller GPUs ist. Dieser lineare Anstieg wird bis 200 GByte/s Gesamtbandbreite verlangt, nicht jedoch darüber hinaus. Diese Bandbreiten für jeden einzelnen Knoten sowie über alle Knoten gleichzeitig ist nachzuweisen mittels ior-easy-read aus den IO500-Benchmarks. Die Bandbreite des Speichersystems muss für den schreibenden parallelen I/O pro GPU-Knoten bei einzelner Messung mit nur einem Knoten mindestens eine Bandbreite von G * 25 GBit/s erreichen, wobei G die Anzahl der GPUs pro Knoten ist. Für alle GPU-Knoten gleichzeitig müssen mindestens 90% von K * 25 GBit/s erreicht werden, wobei K die Anzahl aller GPUs ist. Dieser lineare Anstieg wird bis 100 GByte/s Gesamtbandbreite verlangt, nicht jedoch darüber hinaus. Diese Bandbreite für jeden einzelnen Knoten sowie über alle Knoten gleichzeitig ist nachzuweisen mittels ior-easy-write aus den IO500-Benchmarks. Weiterhin wird eine IOP/s-Rate für 4k random read von allen GPU-Knoten aus von mindestens 3 Millionen IOP/s erwartet. Die IOP/s Rate ist nachzuweisen mittel ior-4K-rnd-read aus der IO500-Benchmarks. Als weiterer Abnahmetest für das Speichersystem ist ein vollständiger Lauf der IO500-Benchmarks im extended mode zu erbringen. Alle Ausgabedateien sowie Job-Skripte und Konfigurationsdateien für den Benchmark sind als Teil der Abnahmedokumentation beizulegen. Alle Ausgabedateien sowie Job-Skripte und Konfigurationsdateien für die Benchmarks sind als Teil der Abnahmedokumentation beizulegen. Alle Abnahmetests zu diesen Kriterium sind vor Erklärung der Betriebsbereitschaft durchzuführen.		
KHG D	**Netzwerke**		**200,00 GP**
A 19	**HPC-Interconnect** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Die zu liefernden GPU-, Service-, Storage-, und Management-Knoten sind in einer		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	gemeinsamen HPC-Interconnect-Insel zu organisieren. Das resultierende Netz muss ein funktionierendes und effizientes Routing erlauben, wozu eine balancierte Nutzung paralleler Links gehört. Die Konfiguration des Routings ist Teil des Lieferumfangs. Das angebotene System ist in die existierende InfiniBand-Infrastruktur zu integrieren. Das bestehende InfiniBand-Netzwerk ist aktuell in einem Fat-Tree organisiert mit fünf Top-Level Switchen (L3). Dort (Raum E12, Stellfläche 5, obere Reihe) stehen jeweils 10 HDR200-Ports zur Verfügung, von denen max. 5 für diese Installation geplant sind. Jeder Uplink muss dabei eine Bandbreite von 200 Gbit/s aufweisen. Mit der Integration der neuen Hardware muss die Fat-Tree-Topologie erhalten bleiben. Auch die Server für den Storage, Service- und Management-Knoten sind in diese Insel geeignet zu integrieren. Dabei ist zu beachten, dass alle luftgekühlten Systeme im Nachbarraum E15 hinter einem Brandschott betrieben werden, die Anzahl der durchgeführten Kabel sollten entsprechend der Bandbreitenanforderungen minimiert werden. Eine detaillierte Dokumentation dieses Netzwerks und der Integration in das bestehende InfiniBand-Netz ist Teil des Angebots.		
B 20	**HPC-Interconnect: Integration** Falls kein InfiniBand eingesetzt wird: Es ist durch den Auftragnehmer sicherzustellen, dass alle Knoten netzwerktechnisch an das bestehende InfiniBand-Netzwerk angeschlossen werden und performanten Zugriff auf die bestehenden Lustre-Dateisysteme haben. Die Lösung ist vom Bieter zu Beschreiben. Der Einsatz von InfiniBand als HPC-Interconnect wird mit 10 Punkten bewertet, andernfalls werden 0 Punkte vergeben.		200 GP
A 21	**HPC-Interconnect: Zugriffskontrolle** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) In der bestehenden InfiniBand-Infrastruktur ist die Partition-ID 0x0001 für das IO-		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Netzwerk vergeben. Alle Filesystem-Server haben dabei volle Mitgliedschaft, alle Klienten eingeschränkte. Dies ermöglicht, dass Compute-Hardware temporär mit eingeschränkten Filesystem-Zugriffsrechten (z.B. von Dritten) betrieben werden kann. Wird InfiniBand als HPC-Interconnect angeboten: Die InfiniBand-HCAs der Speichersystem-Server erhalten die IB-Partition-ID 0x8001. Für den Zugriff auf das Speichersystem wird den GPU- und CPU-Knoten die Partition-ID 0x0001 zugeordnet. Eine weitere PID 0x8002 wird den GPU- und CPU-Knoten für die Kommunikation untereinander zugeordnet. Die technischen Lösungen zur Ausweitung der Zugriffskontrolle auf die Systeme müssen im Angebot dokumentiert werden. Wird nicht InfiniBand als HPC-Interconnect angeboten: Es ist sicherzustellen, dass alle - und ausschließlich - Knoten der Bestandssysteme mit der PID 0x0001 Zugang zum Speichersystem bekommen. Im Angebot ist zu dokumentieren, wie ausgewählten GPU-Knoten der Zugriff auf das Speichersystem und auf die bestehenden Filesystem-Server entzogen werden kann.		
A 22	**Administrations- und Service- Netzwerk** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Als Administrations- und Service-Netzwerk soll für alle Knoten ein Netzwerk auf Ethernet-Basis mit einer Übertragungsrate von mindestens 1 Gbit/s bereitgestellt werden. Aus Gründen der Wartbarkeit und der einfachen Integration in die vorhandene Netzwerk-Managementinfrastruktur sollten bevorzugt Switche mit dem Netzwerkbetriebssystem Junos eingesetzt werden. Ausnahmen sollten im Angebot begründet werden. Dieses Netzwerk soll - in separaten VLANS - die Steuerung der gelieferten Hardware (IPMI, SNMP) ermöglichen. Insbesondere soll auch die Konsolenweiterleitung aller Knoten über dieses Netz eingerichtet werden. Soweit möglich und technisch sinnvoll sollen separate BMC-Links zu den Knoten vermieden werden.		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Für gemanagte Switche werden separate Management-Links außerhalb dieses Netzwerks vom AG bereitgestellt. Für die Uplinks der Switche und als Verbindung zwischen den Räumen stehen im Raum E12 und im Raum E15 jeweils 2x6 25GbE-Ports zur Verfügung. Eine detaillierte Dokumentation dieses Netzwerks ist Teil des Angebots.		
A 23	**Anbindung existierender HPC-Dateisysteme** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Am ZIH sind Lustre-Filesysteme (DDN Exascaler 6) im Einsatz. Diese müssen auf allen Knoten nativ gemountet werden. Beim Auftraggeber vorhandene Kernel-Module können benutzt werden. Alternativ sind die Module vom Auftragnehmer bereitzustellen und auf Anfrage zu aktualisieren, um beispielsweise Kernel-Updates durchführen zu können.		
A 24	**HPC-Interconnect: Abnahmetests** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Es sind Bandbreitentests zwischen zwei Knoten am selben Leaf-Switch sowie zwischen zwei Knoten an verschiedenen Switches (ggf. Leaf-Switches) zu demonstrieren und zu dokumentieren. Sie müssen die nominelle Bandbreite bis auf 5% erreichen, wenn keine andere Kommunikation im HPC-Interconnect stattfindet. Es ist das Erreichen der Uplink-Bandbreite aus der GPU-Insel bis auf 5% der nominellen Bandbreite zu demonstrieren und zu dokumentieren. Bieter können geeignete Benchmarks vorschlagen und verwenden, sofern der AG sie nachvollziehen und selbst wiederholen kann. Empfehlung des AG sind Messungen mit den Kommandos ib_write_bw oder ib_read_bw.		
KHG E	**Software und Management-Knoten**		**0,00 GP**
A 25	**Cluster-Betriebsinfrastruktur** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium)		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Für die Verwaltung der HPC-Cluster sind beim Auftraggeber ein Confluent-Server und eine BlueBanquise-Infrastruktur in Betrieb. In enger Zusammenarbeit mit dem Auftraggeber integriert der Auftragnehmer die neue Hardware in die vorhandene Infrastruktur. Für Managementaufgaben, wie Monitoring, Logging und Alerting sind (mindestens) zwei luftgekühlte Management-Knoten mit redundanten Netzteilen vorzusehen und einzurichten. Die Management-Knoten werden im Raum E15 (Anlage 03 "Stellplan", Stellfläche 7) installiert. Sie sollen je mindestens 32x86_64-Kerne, 512 GB RAM und 8 TB lokale SSD aufweisen. Alle Disks sollen paarweise in redundanten RAIDs laufen. Diese Knoten sind in das HPC-Interconnect integriert. Der Auftragnehmer richtet in enger Absprache mit dem Auftraggeber ein Alerting ein, so dass typische Ausfälle der gelieferten Hardware (wie etwa DIMM-Error, Fehler im HPC-Interconnect, Throttling, Plattenausfall) erkannt und automatisch an das Ticketsystem des Auftragnehmers gemeldet werden. Eventuell darüber hinaus gehende Telemetriedaten aus dem Cluster oder dem Speichersystem dürfen den Cluster nicht verlassen. Über geeignete Container oder Virtuelle Maschinen, die auf den Managementknoten laufen, sind diese Daten lokal zu sammeln. Darüber kann dann der Auftragnehmer Zugang zu diesen Informationen erhalten. Alle betriebskritischen Dienste müssen mit hoher Verfügbarkeit (HA) über die Management-Knoten hinweg vom AN konfiguriert werden. Der AN stellt ausreichend leistungsfähige Hardware bereit für die Erfassung von Job-bezogenen Zeitreihen aus dem Dateisystem und deren Speicherung über 60 Tage. Als Betriebssystem für alle Knoten (GPU+Management) soll RockyLinux/RHEL 9 zum Einsatz kommen. Die Wahl einer anderen Linux-Distribution muss im Angebot angemessen und nachvollziehbar begründet werden. Falls sich Abhängigkeiten zwischen Linux-Kernel, Dateisystem-Software und dem Software-		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Stack des Netzwerks ergeben, ist der AN dafür zuständig, jeweils passende Software-Kombinationen über die gesamte Laufzeit des Systems im Rahmen von Wartung und Service bereitzustellen. Falls die Sicherheit oder Stabilität des Systems nur mit aktualisierter Software erreichbar ist, müssen entsprechende Pakete zeitnah bereitgestellt werden. Veröffentlichte sicherheitsrelevante Schwachstellen des Betriebssystems sollten nach Möglichkeit durch sofortige Aktualisierung behoben werden. Die entsprechenden Updates werden durch den AN umgehend (innerhalb eines Werktages) in dafür vorgesehenen Repositories bereitgestellt. Eine strikte Bindung an bestimmte Kernel-Versionen besteht nicht. Die gelieferte Umgebung muss es ermöglichen, auf dem System selbst compilierte Software nach Auswahl des AG auszuführen. Die NUMA-Domains der CPUs müssen optimal an die tatsächliche Hardware angepasst werden.		
A 26	**Slurm** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Der AN liefert in Abstimmung mit dem ZIH initiale Slurm-Konfigurationen für nicht-exclusive Jobs auf den Knoten. Darin müssen insbesondere • die Isolation von Jobs (cgroups) und • die Zuordnung von SMT-Threads (pinning) und GPUs (gres.conf) zu den durch Slurm verteilten Kernen mit dem Ziel, dass Jobs topologisch benachbarte GPUs und CPUs erhalten, berücksichtigt werden. Für den Betrieb von Slurm-Controler und Datenbank stellt der AG einen Container bereit.		
A 27	**Sicherheit** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Es ist ein Konzept für die Sicherheit des Clusters zu erarbeiten und bereits als Teil des Angebots einzureichen. Grundlegende Punkte darin sind beispielsweise: • Betrieb mit aktivem SELinux, • Ersetzen von Standardpasswörtern durch nichttriviale Passwörter, • Konfiguration von Firewalls auf den Knoten,		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	• Benutzung privater VLANs zur Port-Isolation, so dass Knoten und Infrastruktur maximal geschützt werden. Die Umsetzung des Konzepts ist Voraussetzung für die Erklärung der Betriebsbereitschaft.		
A 28	**MPI** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Die im Rahmen der Benchmarks benutzte MPI-Variante ist mitzuliefern und zusammen mit dem restlichen Software-Stack während der gesamten Laufzeit des Clusters zu aktualisieren, entsprechend der Release-Zyklen dieser MPI-Variante. Der Support muss auch Hardware-nahe MPI-Probleme abdecken.		
KHG F	**Kühlung und Energieeffizienz**		**1.600,00 GP**
A 29	**Warmwasserkühlung für GPU/ Service-Knoten** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Es ist eine direkte Warmwasserkühlung (DLC) für • alle GPU-Knoten aus KHG A und • bevorzugt für alle Service-Knoten aus KHG B (siehe B14 "Rack-Integration") vorzusehen. Die Warmwasserkühlung für diese Knoten soll unter repräsentativen, hohen Arbeitslasten einen Anteil von P >= 90% der Abwärme abführen und eine Temperaturspreizung von mindestens 13 K erreichen. P ergibt sich dabei als der Quotient der ins Warmwasser abgegebenen Abwärme und der zugeführten elektrischen Energie jeweils für alle im Raum E12 installierten Knoten und Komponenten. Beides muss als Teil von A37 "Abnahmetests für Kühlung und Energieeffizienz" nachgewiesen werden. Die o.g. Werte und alle weiteren Details zu Temperaturen, Differenzdrücken, Volumenströmen, Wasserqualität etc. sind Anlage 05: "Betriebskonzept des Lehmann-Zentrum - Rechenzentrum (LZR) der TU Dresden, Anhang B Vorgaben zur Beschaffung/Installation von IT-Systemen" zu entnehmen. Die Installation erfolgt an		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	"Kälte 2 (Warmwasser)", ein zusätzlicher Kaltwasser-Kreislauf ist nicht vorgesehen. Es ist eine Kühlung mit ausschließlich Warmwasserkühlung mit 35 °C Vorlauftemperatur vorgesehen. Bitte beachten Sie besonders die Anforderung, dass die Wasserqualität durch installierte Komponenten nicht gefährdet werden darf, es dürfen nur fabrikneue, nicht verunreinigte Komponenten zum Einsatz kommen und es darf keine Druckprüfung o.ä. mit Wasser/Glykol-Gemischen erfolgt sein. Der Anteil P wird vorzugsweise an einem repräsentativen Rack ermittelt, indem durch den AG die elektrische Leistungsaufnahme und die thermische Leistung bestimmt werden. Die Messung erfolgt bei voller Last. Das System muss im Produktivbetrieb weiterhin die Möglichkeit bieten, Messungen von den vorhandenen Sensoren in den Warmwasserkühlungskomponenten (Cooling Distribution Units, Side Cooler, usw.) maschinenlesbar bereitzustellen. Die Aktualisierung der Messwerte muss mit mindestens einem Sample pro dreißig Sekunden erfolgen. Vom AN ist eine genaue Dokumentation der verfügbaren Messpunkte der Warmwasserkühlung bereitzustellen. Diese beinhaltet mindestens die Aspekte: • Auflistung verfügbarer Messpunkte • Spezifikation der Komponenten der Warmwasserkühlung und Position der Messpunkte innerhalb der Warmwasserkühlung • Dokumentation der Schnittstelle zum Auslesen der Messwerte • Spezifikation ob Messpunkte von Sensoren gemessen sind oder berechnet werden		
B 30	**Effizienz und Messungen der Warmwasserkühlung** Für dieses Kriterium werden max. 10 Leistungspunkte vergeben für die Effizienz der Warmwasserkühlung (bis zu 7 LP) und für die Eigenschaften der Messung. Für jeden ganzen Prozentpunkt den der Quotient P aus Kriterium A29 "Warmwasserkühlung für GPU/ Service-Knoten" über 90% liegt, gibt es einen Leistungspunkt, maximal bis zu 97%		500 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	beziehungsweise sieben Leistungspunkte. Bitte geben Sie den Anteil P und die Spreizung an. Es werden weitere drei Leistungspunkte anhand folgender Leistungen für Messungen an den Warmwasserkühlungskomponenten vergeben: (A) Ein Leistungspunkt wird vergeben, wenn die bereitgestellten Messungen im Produktivbetrieb es erlauben, die über die Warmwasserkühlung abgeführte Abwärme für alle Racks zu bestimmen, insbesondere sollen die Messungen für alle dafür notwendigen Sensoren bereitgestellt werden. (B) Ein weiterer Leistungspunkt wird vergeben, wenn (A) erfüllt ist und das Aktualisierungsintervall der Messungen kleiner oder gleich zehn Sekunden ist. (C) Ein weiterer Leistungspunkt wird vergeben, wenn (A) erfüllt ist und für alle Messpunkte mindestens folgende Auflösungen erreicht werden: • 0,1 K für Temperaturmessungen • 100 W für elektrische und thermische Leistungsmessungen • 0,1 kWh für Energiemessungen • 0,25 m³/h bei Durchflussmessungen • 0,1 bar bei Druckmessungen		
A 31	**Luftkühlung für Storage-Knoten** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Für die Storage-Hardware und die Management-Knoten ist Luftkühlung im Serverraum E15 in vorhandenen Racks im Warmgangcluster vorgesehen. Es sind die Vorgaben und Details in Anlage 05 "Betriebskonzept des Lehmann-Zentrum - Rechenzentrum (LZR) der TU Dresden, Anhang B Vorgaben zur Beschaffung/ Installation von IT-Systemen" maßgeblich. Dies gilt gleichermaßen für Service-Knoten, falls diese mit Luftkühlung angeboten werden.		
A 32	**Energiespar-Möglichkeiten in Hardware und Software** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Der Auftraggeber legt großen Wert auf		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	einen energieeffizienten Betrieb aller Rechner- und Speicherressourcen. Alle nach Stand der Technik üblichen Stromsparmechanismen aller Komponenten (CPUs, GPUs, weitere) müssen standardmäßig aktiv sein. Insbesondere müssen auf sämtlichen Komponenten, die keinem aktiven Job zugeordnet sind, alle verfügbaren Schlafzustände aktiv sein. Die Verwendung von Standby ist jedoch nicht notwendig.		
A 33	**Basis-Energie-Messungen** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Das System (GPU- und Service-Knoten, KHG A und KHG B) muss im Produktivbetrieb die Möglichkeit bieten, Messungen der elektrischen Wirkleistungsaufnahme nach der Green500 "Energy Efficient High Performance Computing Power Measurement Methodology" (nachfolgend Green500 PMM genannt) mindestens mit Level 1 durchzuführen. Details siehe https://www.top500.org/static/media/uploads/methodology-2.0rc1.pdf oder die jeweils aktuellen Version. Dies bedeutet insbesondere: • Aktualisierung der Messwerte mit mindestens 1 Sample pro Sekunde • spezifizierte Genauigkeit der Messungen mit 5% oder besser • Messung eingangsseitig der Leistungsumwandlung • Messung oder Spezifikation des Stromverbrauchs der Netzwerkkomponenten außerhalb der Rechenknoten Ergänzend zu den Anforderungen von "Level 1" muss die Messung mit der Granularität einzelner Rechenknoten möglich sein und sämtliche Rechenknoten abdecken. Das Auslesen der Messdaten zur Leistungsaufnahme muss skalierbar und zuverlässig von außerhalb der Rechenknoten möglich sein und darf dabei keinen Overhead auf den Rechenknoten erzeugen. Vom AN ist eine genaue Dokumentation der verfügbaren Messpunkte bereitzustellen. Diese beinhaltet mindestens die Aspekte: • Auflistung verfügbarer Messpunkte • Spezifikation der Komponenten des		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Stromnetzes und Position der Messpunkte innerhalb des Stromnetzes • Dokumentation der Schnittstelle zum out-of-band Auslesen der Messwerte • (Wenn vorhanden) Dokumentation der Schnittstelle zum inband Auslesen der Messwerte • Spezifikation der Genauigkeit der Messungen • Spezifikation der Rate, mit der sich über die Schnittstelle verfügbare Werte ändern • Spezifikation der internen Abtastrate		
B 34	**Erweiterte Energie-Messungen** Für einen sehr effizienten Betrieb sind auch detaillierte Energie-Monitoring-Möglichkeiten wichtig, sowohl um über Policies oder Accounting eine energiesparende Nutzung belohnen zu können als auch um Energieeinsparungsmaßnahmen gezielt steuern und evaluieren zu können. Daher werden entsprechende Monitoring-Einrichtungen, die über den üblichen Stand der Technik hinausgehen, mit maximal 10 Leistungspunkten bewertet. Für die vier folgenden Eigenschaften (A bis D) der Energie-Messungen werden unabhängig voneinander Punkte vergeben. A) Bessere Genauigkeit (bis zu 3 Punkte): 2 Punkte, wenn das System im Betrieb Messungen nach Green500 PMM Level 2 ermöglicht und diese Messungen getrennt pro GPU-Knoten für alle GPU-Knoten möglich sind. 3 Punkte, wenn das System im Betrieb Messungen nach Green500 PMM Level 3 ermöglicht und diese Messungen getrennt pro GPU-Knoten für alle Rechenknoten möglich sind. B) Bessere zeitliche Auflösung (bis zu 3 Punkte): 2 Punkte, wenn Leistungsmessungen auf Knotenebene mit einer Ausleserate von mindestens 10 aktualisierten Werten pro Sekunde möglich sind. Dabei werden nur physikalische Messungen, die den kompletten GPU-Knoten abdecken berücksichtigt. Das Auslesen der Messdaten muss für Nutzer:innen (nicht nur Admins) möglich sein und darf keinen erheblichen Overhead erzeugen (max. 5% Performanceverlust einer synchronisierten parallelen Anwendung). Es ist zulässig, dass die Werte und zugehörige Zeitstempel		200 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	zwischengespeichert und gesammelt ausgelesen werden. 3 Punkte, wenn alle vorgenannten Anforderungen für "2 Punkte" erfüllt sind, die Leistungsmessungen auf Knotenebene jedoch mit einer Ausleserate von mindestens 100 aktualisierten Werten pro Sekunde möglich sind. C) Messung einzelner Komponenten (2 Punkte): Wenn Leistungsmessungen innerhalb des Knotens, mindestens auf der Ebene einzelner CPU-Sockets möglich sind, wird ein Punkt vergeben. Sind zudem Leistungsmessungen innerhalb des Knotens mindestens auf der Ebene einzelner GPUs möglich, wird ein weiterer Punkt vergeben. Es werden dabei jeweils nur physikalische Messungen berücksichtigt. D) Job-Energieaccounting (1 Punkt): Wenn es für reguläre Nutzer:innen möglich ist, im SLURM Accounting für exklusive Jobs, den Energieverbrauch aller beteiligten GPU-Knoten zu ermitteln (SLURM ConsumedEnergy, ConsumedEnergyRaw). Falls für die GPU-Knoten wenigstens ein Punkt in einer der Eigenschaften A bis C vergeben wurde, wird ein weiterer Punkt vergeben werden, sofern folgende Eigenschaft erfüllt ist: E) Erweiterte Energie-Messungen für Service-Knoten (1 Punkt): Alle Kriterien der Eigenschaften A bis C, die von den GPU-Knoten erfüllt werden, werden auch für die Service-Knoten erfüllt. Für die Eigenschaft C entfällt die Forderung der Leistungsmessung für einzelne GPUs entsprechend. Bitte beachten Sie, dass entgegen der Angaben in https://www.top500.org/static/media/uploads/methodology-2.0rc1.pdf unter "Table 3.1, 1a) Granularity", das interne Sampling für Level 3 mit wenigstens 120 Hz für Wechselstrom (AC) und 5 kHz für Gleichstrom (DC) zu erfolgen hat.		
B 35	**Energieeffizienz des HPL-MxP-Benchmark** Die Energieeffizienz des HPL-MxP-Benchmark-Laufs aus B8 "Benchmark-Leistung mit HPL-MxP" (identischer		600 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Benchmark-Lauf, ein separater, Energie-optimierter Lauf ist nicht zulässig) wird ermittelt und nach der Vorgabe der Green500 PMM (mind. Level 1) mit GFlops/Watt bewertet. Es sind bis zu zehn Bewertungspunkte erreichbar wie folgt: 1 Punkt für >= 600 GFlops/W 2 Punkte für >= 700 GFlops/W 3 Punkte für >= 800 GFlops/W 4 Punkte für >= 900 GFlops/W 5 Punkte für >= 1000 GFlops/W 6 Punkte für >= 1100 GFlops/W 7 Punkte für >= 1200 GFlops/W 8 Punkte für >= 1300 GFlops/W 9 Punkte für >= 1400 GFlops/W 10 Punkte für >= 1500 GFlops/W Falls die angebotenen GPUs einen gültigen Lauf des HPL-Benchmarks (https://top500.org/project/linpack/) zulassen, kann der AG verlangen, dass der AN diesen Benchmark zur Abnahme in Energie-optimierter Konfiguration durchführt und die Konfiguration angibt. Wertungsrelevant für dieses Kriterium ist jedoch nur der HPL-MxP-Benchmark.		
B 36	**Energieeffizienz im Idle-Modus** Der Energieverbrauch im Idle-Modus soll möglichst gering sein. Unter Idle-Modus ist hierbei zu verstehen, dass das System bereit ist, Jobs zu verarbeiten, jedoch derzeit kein Job ausgeführt wird oder einen "sleep"-Job ausführt. Tiefe ACPI S- und G-States sind nicht erlaubt, tiefe C-States sind hingegen erlaubt und gewünscht. Die Bewertung erfolgt anhand des Quotienten aus HPL-MxP-Gesamtleistung (B8) geteilt durch Idle-Leistungsaufnahme des Gesamtsystems mit einer Messung entsprechend der Vorgabe der Green500 PMM. Es sind bis zu zehn Bewertungspunkte erreichbar wie folgt: 1 Punkt für >= 2100 GFlops/Idle-Watt 2 Punkte für >= 2700 GFlops/Idle-Watt 3 Punkte für >= 3300 GFlops/Idle-Watt 4 Punkte für >= 3900 GFlops/Idle-Watt 5 Punkte für >= 4500 GFlops/Idle-Watt 6 Punkte für >= 5100 GFlops/Idle-Watt		300 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	7 Punkte für >= 5700 GFlops/Idle-Watt 8 Punkte für >= 6300 GFlops/Idle-Watt 9 Punkte für >= 6900 GFlops/Idle-Watt 10 Punkte für >= 7500 GFlops/Idle-Watt		
A 37	**Abnahmetests für Kühlung und Energieeffizienz** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Es ist zu demonstrieren, dass unter dauerhaft hoher Last ein stabiler Betrieb möglich ist und dass die Vorgaben aus A29 "Warmwasserkühlung für GPU/ Service-Knoten" sowie die Zusagen aus B30 "Effizienz und Messungen der Warmwasserkühlung" eingehalten werden. Diese Demonstration muss mit mindestens 1h Betrieb unter durchgängig hoher Last auf allen im Raum E12 installierten Knoten erfolgen. Die entsprechende Last kann mit dem HPL-Benchmark, HPL-MxP oder anderen, zwischen AN und AG vereinbarten Benchmarks erzeugt werden. Dieser Test ist vor Erklärung der Betriebsbereitschaft nachzuweisen. Der stabile Betrieb und die Vorgaben nach oben genannten Kriterien müssen auch gewährleistet sein, wenn der AG Tests mit maximal möglicher Rechenlast durchführt, das heißt mit Codes, die maximalen Strom- und Kühlungsbedarf erzeugen.		
KHG G	**Installation und Wartung**		**0,00 GP**
A 38	**Installation** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) In Abstimmung mit dem AG sind die Systeme im LZR-Gebäude der TU Dresden zu installieren und alle Strom- und Netzwerkverbindungen herzustellen. Eine funktionsfähige Grundkonfiguration aller Systeme ist herzustellen. Der AN hat sich dazu mit dem ZIH abzustimmen. Alle angebotenen Systeme sind für einen Serverbetrieb von 365x24 Stunden pro Jahr ausgelegt und zu mindestens 99% pro Monat verfügbar. 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.		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Die Namen und IP-Adressen aller Systeme werden vom AG festgelegt. Um den reibungslosen Betrieb aller Systeme zu gewährleisten, ist die Pflege einer Systemdokumentation erforderlich. Das ZIH betreibt zu diesem Zweck eine CMDB (Configuration Management Database). Es wird erwartet, dass der AN für das zu liefernde System den jeweils aktuellen Hardware-Bestand, Hardware-Adressen, IP-Adressen, Netzwerkverbindungen usw. über die Laufzeit des Service-Vertrages pflegt. Dazu sind alle entsprechenden Informationen initial als xls-Tabelle mit vom ZIH vorgegebener Struktur bereitzustellen. Bei späteren Änderungen (z.B. Austausch von Knoten, Erweiterung) ist diese Tabelle zu aktualisieren. In Absprache mit dem ZIH können auch andere Wege gesucht werden, um die Datenbank direkt zu aktualisieren, etwa direkt über mysql. Ausführliche Einweisungen und Schulungen der Administratoren des AG in angemessenem Umfang sind mit anzubieten.		
A 39	**Elektrischer Anschluss** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Die mitzuliefernden PDUs für die Compute-Knoten müssen folgende Anforderungen erfüllen: Es ist nur eine A-Versorgung vorgesehen (ohne Redundanz). Auf redundante Netzteile kann verzichtet werden. 2n-redundante Netzteile sind per Y-Kabel an eine PDU-Buchse anzuschließen. Ziel sind möglichst wenige PDUs pro Rack mit: • Eingang: 63A bis 80A (nach Rücksprache bis max. 100A), 3L+N+PE • Kabellänge mind. 5m, nicht schaltbar, 25mm² Querschnitt • bevorzugt ohne integrierte Elektronik, • geeignet für den dauerhaften Betrieb bei mind. 50°C • Pro Phase mind. 4×16A Sicherung, bevorzugt nur C19 Buchsen Gleichwertige Lösungen für die PDUs sind prinzipiell möglich, wenn sie vollumfänglich die gleiche Funktionalität bieten und die Kompatibilität mit der Rechnerraum-		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Infrastruktur sichergestellt ist. Vor Inbetriebnahme muss die Prüfung der elektrischen Anlagen nach DIN VDE 0100-600 durch den AN als Leistungsbestandteil durchgeführt werden. Die entsprechenden Prüfprotokolle sind dem AG zur Verfügung zu stellen. Bezüglich der Installation und für den Betrieb sind Anlage 04 "Betriebskonzept des Lehmann-Zentrum - Rechenzentrum (LZR) der TU Dresden, Anhang A: "Anlieferung und Materialtransport" und Anlage 05: "Betriebskonzept des Lehmann-Zentrum -Rechenzentrum (LZR) der TU Dresden, Anhang B Vorgaben zur Beschaffung/ Installation von IT-Systemen" maßgeblich.		
A 40	**Netzwerk-Verkabelungen** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Alle Systeme verfügen über eine Schnittstelle für Remote Management via IPMI 2.1 oder neuer. Alle erforderlichen Lizenzen sind mitzuliefern. Für jede Komponente ist ein Link zu einem mitzuliefernden Ethernet-basierenden Management-Netzwerk herzustellen. Alle Komponenten sind mit allen vorgesehenen Netzwerk-Arten zu verkabeln. Anzuschließen sind in gleicher Weise alle erforderlichen Netzwerk-Verbindungen zu den existierenden HPC-Komponenten. Die Anbindungen sind mit den nominalen Geschwindigkeiten (theoretische Links-Bandbreite x Anzahl der Links) vorzusehen. Die Bandbreiten zu allen Komponenten der neuen Insel soll ausgewogen sein. Der Auftraggeber ist an vorteilhaften Lösungsvorschlägen interessiert, wird diese im Rahmen der Verhandlungen diskutieren und bewerten und die insgesamt beste und wirtschaftlichste Variante als konkrete Anforderung in das finale Leistungsverzeichnis aufnehmen. Alle notwendigen Netzwerk- und Strom-Kabel sind mitzuliefern und an beiden Enden (identisch) zu beschriften. Auf den Labels muss mindestens stehen: Name der verbundenen Systeme (from/to) und jeweils für diese: Port und Einbauposition (Rackname und HE). Im Zweifelsfall entscheidet nach Absprache der AG über den zu labelnden Text.		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Alle zur Verbindung der Komponenten benötigten Transceiver und Kabel sind mitzuliefern. Kabelfarben sind mit dem AG abzustimmen. Die Konfiguration der Netzwerke muss in Zusammenarbeit mit dem ZIH erfolgen.		
A 41	**Warmwasserkühlung und Racks** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Für die Compute-Knoten mit Warmwasserkühlung sind Racks inkl. der Warmwasser-Kühlungs-Infrastruktur (Sekundärkreis) mitzuliefern. Am Aufstellplatz in der E12 soll ein gebäudeseitig Verrohrung mit Rack-Anschlüssen für max. 2 l/s bevorzugt genutzt werden, dann sind die erforderlichen Cooling Distribution Units (CDUs) in die Racks zu integrieren. Nach Rücksprache und Freigabe durch den AG ist auch die Installation von CDUs im Plenums-Geschoss im Raum S12 möglich. In diesem Fall sind die notwendigen Änderungen der Verrohrung für den Anschluss der CDUs an die gebäudeseitige Verrohrung als Teil des Angebots vom AN durchzuführen. Anlage 05: "Betriebskonzept des Lehmann-Zentrum -Rechenzentrum (LZR) der TU Dresden, Anhang B Vorgaben zur Beschaffung/ Installation von IT-Systemen" ist zu beachten, insbesondere müssen CDUs mit Durchgangsventilen (2-Wege) ausgestattet sein. Sofern Nicht-Standard-Racks (insb. nicht 60cm breit) vorgesehen werden, ist die Aufstellung unbedingt vorab mit dem AG abzustimmen hinsichtlich Geometrie, Last und Anschlüssen. Anlage 04 "Betriebskonzept des Lehmann-Zentrum - Rechenzentrum (LZR) der TU Dresden, Anhang A: Anlieferung und Materialtransport" ist zu beachten. Leicht über die Grenzen hinausgehende Lasten müssen vom AG ausdrücklich freigegeben werden. Datenblätter von Racks und ggf. vorgesehener Grundrahmen sind dem Angebot beizulegen. Bitte geben Sie die Last der Racks inklusive aller Knoten, Netzwerk-Switche, anderen Komponenten und Kühlmedien an. Aussparungen für Durchführung von		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Medien/ELT/KLT durch Doppelbodenplatten sind so klein wie möglich vorzusehen und sind in einem Grundriss einzuzeichnen.		
A 42	**Nachnutzung bestehender Serverracks** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Für die Compute-Knoten mit Warmwasserkühlung plant der Auftraggeber die kostenfreie Nachnutzung von bis zu 5 Serverracks inkl. PDUs anzubieten. Die Serverracks stammen aus einem aktuell noch im Betrieb befindlichen GPU-Cluster und werden voraussichtlich zum 01.05.2026 frei. Die Nachnutzung der aktuell verbauten Sidecooler ist vom AG aus Gründen der Energieeffizienz nicht gewünscht. Die Nachnutzung der aktuell verbauten CDUs ist vom AG nicht gewünscht. Der AN hat vorab die Kompatibilität der vorhandenen Racks mit der geplanten Infrastruktur zu prüfen. Hierfür bietet der AG Vor-Ort-Besichtigungen der aktuellen Infrastuktur des GPU-Clusters an. Die Wartung sowie der Ersatz bei Ausfall der PDUs verbleiben beim AG, der auch die Kosten dafür trägt. Die technische Beschreibung der Serverracks ist wie folgt: • Racks mit Sidecoolern • Die Racks sind wärmeisoliert und praktisch luftdicht verschlossen. Sie können daher prinzipiell als thermisch autarke Systeme betrachtet werden. Die verbauten PDUs entsprechen den Anforderungen aus A39. Im Falle einer Nachnutzung der Serverracks und PDUs können diese voraussichtlich auf Stellfläche 5 verbleiben, so dass keine oder nur kleine Umbaumaßnahmen bzgl. der Abgangskästen und Verrohrung erfolgen müssen. Etwaige Umbaumaßnahmen sind vom AN zu leisten. Eine Aufstellung auf Stellfläche 4 ist zusammen mit dem AG zu prüfen. Die beiden Aufstellflächen sind in Anlage 03 Stellplan gekennzeichnet. Die Konkretisierung der Optionen zur Aufstellfläche kann erst im finalen Leistungsverzeichnis erfolgen.		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	Der AG möchte alle Bieter im Sinne einer ressourcenschonenden Lösung zur Prüfung der Nachnutzung für das zu beschaffende GPU-Cluster ermuntern. Die Erstangebote sollen die Nachnutzung ausdrücklich nicht berücksichtigen. Vom AN erarbeitete Lösungen, die eine Nachnutzung der oben genannten Infrastruktur berücksichtigen, sollen in den Verhandlungsgesprächen skizziert und der Lösung ohne Nachnutzung gegenübergestellt werden. Bei Unvereinbarkeit ist eine Mitteilung mit begründeter Ablehnung vorzulegen.		
A 43	**Brandschutz und Brandmeldung** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Für Racks mit (teilweise) hermetischer Abdichtung (z.B. für thermische Isolation) ist die Wirkung der Stickstoffschnellabsenkung (Brandschutz) nicht gewährleistet. Hier ist je Rack mind. ein Brandmelder zu installieren und über potenzialfreiem Kontakt nach außen zu führen, so dass zum Sachschutz eine automatische externe Abschaltung der Stromzufuhr erfolgen kann. Siehe auch Anlage 05: "Betriebskonzept des Lehmann-Zentrum - Rechenzentrum (LZR) der TU Dresden, Anhang B Vorgaben zur Beschaffung/ Installation von IT-Systemen".		
A 44	**Gewährleistung und Service** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Eine Gewährleistungserweiterung auf mindestens drei Jahre für alle Komponenten ist einzuschließen. Die Umsetzung und Verfolgung von Hardware-Garantieansprüchen wird über den Auftragnehmer abgewickelt. Er ist verantwortlich für die Organisation und Integration der Ersatzkomponenten in das Cluster-System. Der Auftragnehmer verpflichtet sich, den ursprünglichen Zustand des Systems wieder herzustellen. Garantieansprüche bleiben davon unberührt. Bei Bekanntwerden von Konstruktionsfehlern kann der Auftraggeber auf den Austausch der Komponenten auf Kosten des Herstellers bestehen. Es wird ein eingeschlossener Hard- und Software-Service (Wartung/Reparatur) inkl. aller Teile, Updates und Upgrades für mindestens drei Jahre gefordert und Anschlussverträge für das 4. und 5. Jahr		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	(einzeln). Einschränkungen bezüglich des Zeitpunktes für den Abschluss dieser Anschlussverträge sind darzulegen. Ein umfassender Hard- und Software-Support an Werktagen während der Hauptbetriebszeiten ist anzubieten. Als Wartungslevel soll vereinbart werden: • Betriebszeit: 5x9 (Mo. - Fr. 8:00-17:00 Uhr) • Die Reaktionszeiten richten sich nach Nummer 5.1.1.2 des EVB-IT Systemvertrages • Wiederherstellungszeit: angestrebt sind 2-3 Werktage, maximal 4 Werktage Wenn über 5% der Rechenknoten als Hardwarefehler gemeldet sind, so erwartet der Auftraggeber einen Austausch oder eine Reparatur dieser Knoten innerhalb von 24 Stunden. Die Bestimmungen nach Ziffer 4.1.2 EVB- IT System-AGB sind zu beachten. Die gesamte Hardwarewartung wird vom Auftragnehmer durchgeführt. Für die Vertragslaufzeit ist der Einsatz von Personal vor Ort für den ggf. erforderlichen Austausch von Hardware-Komponenten zu leisten. Arbeitsplätze können bei Bedarf bereitgestellt werden. Die Verfügbarkeit von Ersatzteilen aus identischen oder gleichwertigen kompatiblen Komponenten ist für fünf Jahre ab Inbetriebnahme zu garantieren. Eine ausreichende Anzahl von Ersatzteilen bei häufiger ausfallenden Komponenten (Disks, DIMMs), wenn möglich auch Ersatzknoten, ist vor Ort zu lagern. Reparierte bzw. ausgetauschte Komponenten müssen die gleichen BIOS/ Firmware-Einstellungen und -Versionen haben, wie die anderen installierten Systeme gleicher Art. Der Systemsupport schließt folgende Aktivitäten ein, die vom Auftragnehmer in Absprache mit den Systemadministratoren zu erbringen sind: • Firmware-, Microcode- und Treiber-Update-Service für die Laufzeit der Systeme; Prüfung und Durchführung mindestens einmal jährlich im Rahmen der Systemwartung		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	• Installation, Support und Updates des Betriebssystems und aller zum Betrieb erforderlichen Software-Komponenten durch den Auftragnehmer • Fehlersuche bei Systemabstürzen • Tuning von GPUs und Betriebssystem mit Hinblick auf System-Performance und Enerieeffizienz (initial und bei Problemfällen) • Installation, Tuning und Betrieb inkl. Wartung des Dateisystems • Unterstützung bei der Konfiguration des Batch-Systems • Konfiguration und Tuning der System-Management-Software (initial und bei Problemfällen) • Unterstützung bei der Analyse und Behebung hardwarenaher Performance-Probleme Eine Vor-Ort-Instandsetzung der Systeme im Störungsfall sowie deutschsprachiger technischer Support per Telefon und Email für alle Hard- und Software-Komponenten und die Ersatzteile, Anfahrt und Arbeitszeit müssen in der Supportleistung enthalten sein. Updates für alle gelieferten Software-Produkte sind mindestens alle 6 Monate vorzusehen, Sicherheits-Updates sind zeitnah einzuplanen. Im Rahmen der HPC-Strategie des ZIH können zukünftige Erweiterungen des Systems nötig werden. Diese müssen wiederum ausgeschrieben werden. Es ist daher erforderlich, dass der Systemsupport - für die im Rahmen dieser Ausschreibung gelieferten Hard- und Software-Komponenten - weiterhin geleistet wird, auch wenn dem System zukünftig ggf. Komponenten anderer Hersteller hinzugefügt werden. Für den Austausch von leicht zugänglichen Hot-Swap-Komponenten (z. B. Festplatten, Lüfter o.ä.) kann ein Self-Service vereinbart werden (siehe Nummer 12 des EVB-IT Systemvertrages). Bereits mit der Installation muss eine einfach bedienbare Schnittstelle zwischen den Systemadministratoren und dem Auftragnehmer für die Behandlung von Störungen durch ein Ticketsystem für zeitnahe und nachvollziehbare Prozesse sorgen. Der Auftragnehmer hat dafür Sorge zu tragen, dass ausgewählte und gemeinsam mit dem Administratorteam		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	definierte Störungsmeldungen, die automatisch im HPC-System erzeugt werden (z.B. genau klassifizierte Fehler von RAM-Modulen, Disks oder gedrosselte Knoten), auf geeignete Weise automatisiert an das Ticketsystem übergeben und ohne Eingreifen des Auftraggebers bearbeitet werden. Ein Knoten mit Leistungsdrosselung (throttling) infolge von z.B. Problemen mit der Kühlung oder der Spannungsversorgung wird als defekt angesehen. Für das System ist im Betrieb eine mittlere Verfügbarkeit von 99% über jedes Kalender-Quartal zu gewährleisten. Die Verfügbarkeit zu einem bestimmten Zeitpunkt berechnet sich dabei auf folgende Weise: • Eine Unterbrechung des Dateisystems (auch von Teilen) zählt als kompletter Ausfall des Systems (0% verfügbar). • Eine Unterbrechung des Management-Systems zählt als kompletter Ausfall des Systems (0% verfügbar). • Funktionieren 2 oder mehr Service-Knoten nicht komplett, zählt das als kompletter Ausfall des Systems (0% verfügbar). • Sind Dateisystem, Management- und Service-Knoten im obigen Sinne "verfügbar", werden komplett funktionierende GPU-Knoten anteilig im Verhältnis zu den angebotenen GPU-Knoten gewertet. Durch das Mitlaufen von hot-spare-Knoten kann sich auch eine Rate von >100% ergeben, die aber nicht auf andere Quartale angerechnet wird. Kann diese Verfügbarkeit nicht eingehalten werden, gilt im Rahmen der Wartung folgende Regelung: Am Ende jedes Betriebsjahres des Systems kann der AG für die durch unter 99% Verfügbarkeit potenziell entgangenen GPU-Rechenstunden Kompensation verlangen. Die Kompensation erfolgt durch Installation zusätzlicher GPU-Knoten, so dass die entgangen GPU-Rechenstunden in den restlichen Betriebsjahren (nach A3) ab Inbetriebnahme dieser Knoten ausgeglichen werden. Die ggf. nötigen Infrastruktur-Komponenten (Interconnect-Switche, Racks, etc.) sind darin eingeschlossen. Der Auftraggeber		

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	weist auf die Möglichkeit hin, Hot-Spare-Knoten zu installieren, um die geforderte Verfügbarkeit einzuhalten. Im 4. und 5. Jahr des Betriebs kann die Kompensation auch durch andere Leistungen ersetzt werden, die zwischen Auftragnehmer und Auftraggeber verhandelt werden müssen. Vom Bieter wird ein Wartungs- und Supportkonzept vorgelegt, das detailliert auf die in diesem Abschnitt dargelegten Punkte eingeht.		
KHG H	**Abnahme, Projektabschluss, Gesamtbewertung**		**2.000,00 GP**
A 45	**Abnahme und Verfügbarkeitstest** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Im Rahmen der 30-tägigen Abnahmephase wird die Zustand des Systems überwacht und protokolliert. Anlage 06 ("Verfügbarkeit Abnahme") beschreibt detailliert die Protokollierung und die Kriterien zum Bestehen des Verfügbarkeitstests. Während der gesamten Abnahmephase dürfen maximal 10% der insgesamt zur Verfügung stehenden Rechenzeit direkt oder indirekt durch Ausfälle verloren gehen. Während der Bürozeiten im Abnahmezeitraum muss eine Verfügbarkeit von 99% gewährleistet werden.		
A 46	**Projektabschluss und Rechnungslegung** (Ist Ausschlusskriterium) (Ist Ja-oder-Nein-Kriterium) Die Rechnungslegung für die erste Teilrechnung ist bis spätestens 30.11.2025 zu stellen, die Abschlussrechnung spätestens zum 31.10.2026. Die Lieferung muss bis Ende Juni 2026 erfolgen. Die Betriebsbereitschaft muss bis zum 1.9.2026 erklärt werden, so dass die Abnahme Ende September 2026 abgeschlossen werden kann. Der Nutzerbetrieb ist ab 1.10.2026 geplant. Bei Verzögerungen gilt A3.		
B 47	**Gesamtbewertung** Der Auftraggeber nimmt eine Gesamtbewertung des Angebots vor. Eine wichtige Voraussetzung dafür ist eine gute		2.000 GP

Nr.	Bezeichnung	Antwort	Kriteriengewichtung
	und nachvollziehbare Dokumentation der angebotenen Lösungen. Die Bewertung richtet sich auf die Passfähigkeit und Balance der angebotene Hardware für die gewünschten Zwecke. Dabei werden insbesondere die Kombination von Leistungsfähigkeit und Energieeffizienz, die Nutzbarkeit, die Eignung für einschlägige ML- und HPC-Software-Pakete, die Zukunftsfähigkeit und Entwicklungsaussichten der Hardware-Architektur, die Qualität und Verfügbarkeit von Software-Umgebungen des Herstellers sowie von Open Source Communities, den Terminplan für Lieferung und Inbetriebnahme, das Sicherheitskonzept, der Preis und die Leistung der optionalen Positionen sowie die Gesamtdarstellung des Angebots berücksichtigt. Wichtig ist darüber hinaus, ob im Wartungs- und Support-Konzept ausreichend gezeigt wird, wie die geplante hohe Verfügbarkeit im Regelbetrieb erreicht werden kann, um spätere Kompensationen zu vermeiden. Eine balancierte Übererfüllung von Mindestkriterien kann hier positiv bewertet werden, wenn sie dem Einsatzzweck des System dient. Es sind 0 bis 10 Leistungspunkte erreichbar nach Einschätzung des Auftraggebers. Zielerfüllungsgrad: 7-10 Punkte: Die Anforderung ist von großteilig bis vollständig erfüllt. 3-7 Punkte: Die Anforderung ist bedingt erfüllt. 0-3 Punkte: Die Anforderung ist nur sehr bedingt erfüllt.		

Fragebogen 1: Fragen zu KHG A GPU-Compute-Knoten

Fragetitel	Antwort
1.1 Lieferzeit inkl. Herstellung der Betriebsbereitschaft Geben Sie die Lieferzeit und die Zeit zur Herstellung der Betriebsbereitschaft (Abnahmereife) in Kalenderwochen nach Zugang der Bestellung beim Auftragnehmer (Zuschlag) an.	

Fragetitel	Antwort
1.2 Hersteller Geben Sie den Hersteller an.	
1.3 Produktbezeichnung Geben Sie die eindeutigen Produktbezeichnungen sowie den Leistungs- bzw. Lieferumfang mit den Ausstattungsmerkmalen an. Nutzen Sie ggf. eigene Anlagen (z. B. Komponentenliste mit klaren Bezügen zu diesem Dokument), soweit der zur Verfügung stehende Platz nicht ausreicht. Hinweis: Bitte beachten Sie, dass Sie bei Verwendung eigener Anlagen KEINE Änderungen oder Ergänzungen an den Vergabe- und Vertragsunterlagen vornehmen.	
1.4 Anzahl GPU-Knoten Wie viele GPU-Knoten werden insgesamt angeboten?	
1.5 CPUs Welche CPU-Typen werden in den GPU-Knoten verbaut? Wie viele Kerne pro CPU und pro Knoten bieten sie? Wie viele SMT-Threads o.ä. bietet jede CPU und jeder Knoten?	
1.6 GPUs Welche GPU-Typen werden in den GPU-Knoten verbaut und wie viele GPUs pro GPU-Knoten installiert?	
1.7 Rack-Höheneinheiten pro Knoten Wie viele Rack-Höheneinheiten werden pro GPU-Knoten und wie viele pro CPU-Knoten eingenommen?	
1.8 Rack-Höheneinheiten für das System Wie viele Rack-Höheneinheiten werden durch die GPU/ CPU-Knoten insgesamt belegt? Wie viele HE kommen	

Fragetitel	Antwort
für Switches hinzu? Wie viele HE kommen für Kühl-Infrastruktur innerhalb der Racks hinzu? Wie viele HE werden insgesamt inkl. ggf. weiterer Komponenten in den warmwassergekühlten Racks benötigt?	
1.9 TDP Wie hoch sind die TDP der angebotenen CPUs und GPUs?	
1.10 Hauptspeicher Für GPU/CPU-Knoten: welcher DIMM-Typ mit welcher Kapazität wird als Host-RAM verbaut? Wie viele DIMMs pro Host werden verwendet? Falls andere Hauptspeicher-Arten zum Einsatz kommen, mit welcher Anzahl und Kapazität sowie welchen Spezifikationen?	
1.11 Nutzerschulungen Für die Unterstützung der Nutzer bei der Migration ihrer Programme vom bestehenden Capella-System (Nvidia H100) auf den angebotenen Cluster sollen Kurse veranstaltet werden, um zu zeigen, wie die neuen GPUs, lokale NVMes und das parallele Filesystem optimal und skalierbar genutzt werden können. Die Kurse sollen sich sowohl an Einsteiger als auch fortgeschrittene Nutzer richten, bevorzugt mit aufeinander aufbauenden Modulen. Schwerpunkte sind u.a. KI-Frameworks, parallele GPU-Programmierung und Programmierwerkzeuge. Ebenfalls ist ein Workshop "Bring your Code" für die Migration vorzusehen. Beschreiben Sie das Kursangebot und geben Sie den verbindlichen Preis für einen 3-Tages-Kurs (je 8h) im online oder onsite-Format für bis zu 30 Teilnehmende an. Der AG kann nach seiner Wahl einen oder mehrere Kurse zu diesem Preis im ersten Jahr nach Abnahme des Systems beauftragen.	

Fragebogen 2: Fragen zu KHG B Service-Knoten

Fragetitel	Antwort
2.1 CPUs Welche CPU-Typen werden in den Service-Knoten verbaut? Wie viele Kerne pro CPU und pro Knoten bieten sie? Der erreichbare Rmax (HPL) je Knoten ist	

Fragetitel	Antwort
anzugeben.	

Fragebogen 3: Fragen zu KHG C Storage

Fragetitel	Antwort
3.1 Hersteller Geben Sie den/die Hersteller von Software und Hardware des Dateisystems an.	
3.2 Produktbezeichnung Geben Sie die genauen Produktbezeichnungen von Software und Hardware des Dateisystems an.	
3.3 Storage-Devices Geben Sie den Hersteller und Produktbezeichnung der verbauten Storage-Devices an. Geben Sie an, aus welchen NAND-Flash-Typ die Storage-Devices bestehen.	
3.4 Anzahl Storage-Devices Aus wie vielen Storage-Devices besteht das Speichersystem und auf wie viele Knoten sind diese verteilt?	
3.5 Unterliegendes Dateisystem Welches unterliegende Dateisystem wird für die einzelnen Storage-Devices verwendet?	
3.6 Lese- und Schreib-Bandbreiten Wie hoch sind die aggregierten Lese- und Schreib-	

Fragetitel	Antwort
Bandbreiten des Dateisystems?	
3.7 Metadaten-Rate Wie hoch sind IOP/s-Rate für Metadaten-Operationen create, stat, delete, 4k random read?	
3.8 Software-Anforderungen Werden für den Betrieb des parallelen Dateisystems Kernel-Module oder andere spezielle Software-Komponenten benötigt? Welche Einschränkungen gibt es für die Verfügbarkeit des Dateisystem nach Sicherheitsupdates des Linux-Kernels? Geplant ist, bestehende produktive Dateisysteme an den gleichen Knoten anzubinden (aktuell: Lustre 2.14.0-ddn, WEKA 4.4.0). Welche potentiellen Konflikte können sich dadurch ergeben und wie können sie behoben oder umgangen werden? Das Dateisystem soll auch an weiteren Knoten im Bestandssystem (Infiniband) gemountet werden. Gibt es lizenzrechtliche oder technische Bedingungen, die das einschränken?	
3.9 Hardware-Anforderungen Welche dedizierten Ressourcen (z.B. RAM, CPU-Kerne) müssen auf den GPU-Knoten für das Mounten des Dateisystems eingeplant werden?	
3.10 Interfaces Mit welchen Interfaces kann das Speichersystem verwendet werden? (z.B. POSIX, S3, DFS)	
3.11 Datensicherheit Welche Features zur Datensicherheit werden unterstützt (Replication, Snapshots, Ereasure Coding, keine)?	

Fragetitel	Antwort
3.12 Füllstand Bis zu welchen Füllstand des Dateisystems kann die volle Performance garantiert werden? (in Prozent)	

Fragebogen 4: Fragen zu KHG D Interconnect

Fragetitel	Antwort
4.1 Robustheit Gestörte Verbindungen können sich auf das gesamte Netz auswirken. Auf welche Weise soll die Fehlerfreiheit automatisiert gewährleistet werden?	

Fragebogen 5: Fragen zu KHG E Software und Management-Knoten

Fragetitel	Antwort
5.1 Boot-Zeiten Wie lange dauert das Booten aller Compute-Knoten aus dem ausgeschaltetem Zustand?	
5.2 Betriebssysteme Bitte listen Sie unterstütze Betriebssysteme aus der Red-Hat-Familie (9.x) auf. Wie viele Wochen nach Release einer neuen minor Version kann diese vollumfänglich (einschl. Filesystem) für den Betrieb des Clusters genutzt werden?	
5.3 Management-Protokolle für Remote Boot Welche Management-Protokolle für Remote Boot werden unterstützt?	

Fragetitel	Antwort

Fragebogen 6: Fragen zu KHG F Kühlung und Energieeffizienz

Fragetitel	Antwort
6.1 Elektrische Leistungsmessung Welche Möglichkeiten zur elektrischen Leistungs- und Energie-Messung sowie zur Temperatur-Messung bestehen pro Rack, pro Knoten und ggf. für GPUs oder andere Teilkomponenten der Compute-Knoten? Welche zeitliche Auflösungen bieten die Messeinrichtungen? Welche Genauigkeit bieten die jeweiligen Messungen?	

Fragebogen 7: Fragen zu KHG G Installation und Wartung

Fragetitel	Antwort
7.1 Leistungsaufnahme Wie groß ist die elektrische Leistungsaufnahme eines einzelnen GPU-Knotens und CPU-Knotens maximal? Wie ist das Maximum pro Rack-Einschub oder Chassis o.ä.?	
7.2 Stromanschlüsse Über wie viele Stromanschlüsse welcher Art verfügt ein Rack-Einschub?	

Angebot

Mit Unterzeichnung des Angebotes erkennt der Bieter die Forderungen und Angaben des Leistungsverzeichnisses an und bestätigt die Richtigkeit der von ihm gemachten Angaben.	**Nachlass in %:**	
	Gesamtangebotssumme ohne USt. inkl. Nachlass (EUR):	
	Gesamtangebotssumme inkl. USt. und Nachlass (EUR):	