

# Offenes Verfahren

„High Performance Cluster“

Vergabe-Nr.: BAITZ-2024-10

Auftraggeberin:

Martin-Luther-Universität Halle-Wittenberg  
IT-Servicezentrum  
Kurt-Mothes-Straße 1  
06120 Halle



MARTIN-LUTHER-UNIVERSITÄT  
HALLE-WITTENBERG

## Inhaltsverzeichnis

<b>1</b>	<b>Vorbemerkungen .....</b>	<b>3</b>
<b>2</b>	<b>Ziel der Beschaffung .....</b>	<b>5</b>
<b>3</b>	<b>Leistungsbeschreibung .....</b>	<b>6</b>
<b>3.1</b>	<b>Struktur des HPC Clusters .....</b>	<b>6</b>
<b>3.2</b>	<b>Gesamter HPC Cluster .....</b>	<b>7</b>
3.2.1	Kaufmännische Kriterien .....	7
3.2.2	Technische Kriterien .....	7
<b>3.3</b>	<b>Rechenknoten / Compute Nodes .....</b>	<b>10</b>
3.3.1	CPU Compute Nodes .....	10
3.3.2	GPU fp32 Compute Nodes .....	12
3.3.3	GPU fp64 Compute Nodes .....	13
<b>3.4</b>	<b>Zentrale Nodes .....</b>	<b>15</b>
3.4.1	Management Nodes .....	15
3.4.2	Login Nodes .....	17
<b>3.5</b>	<b>Storage .....</b>	<b>18</b>
3.5.1	Storage - Allgemein .....	18
3.5.2	Storage Software .....	18
3.5.3	Storage Hardware .....	19
3.5.3.1	Storage Hardware - Allgemein .....	19
3.5.3.2	Lustre MetaDatenServer und ObjectStorageServer .....	19
3.5.3.3	Lustre MetaDatenTargets und ObjectStorageTargets .....	21
3.5.3.4	NFS Server und NFS Backend Storage .....	22
3.5.3.5	HPC Managementdaten Storage .....	22
<b>3.6</b>	<b>Netzwerk .....</b>	<b>23</b>
3.6.1	Netzwerk - Infiniband .....	23
3.6.2	Netzwerk - Ethernet .....	23
<b>3.7</b>	<b>HPC Software .....</b>	<b>24</b>
3.7.1	HPC Software - Allgemein .....	24
3.7.2	Betriebssystem .....	24
3.7.3	Dienste .....	25
3.7.3.1	Provisionierung .....	25
3.7.3.2	Monitoring .....	26
3.7.3.3	Batchsystem .....	26
3.7.4	Softwareerstellung .....	27
<b>3.8</b>	<b>Schulung und Dienstleistung .....</b>	<b>27</b>
<b>3.9</b>	<b>Vor - Ort Installation und Dokumentation .....</b>	<b>27</b>
<b>4</b>	<b>Ausführungsbestimmungen .....</b>	<b>28</b>
<b>5</b>	<b>Zuschlagskriterien .....</b>	<b>29</b>
<b>5.1</b>	<b>Zuschlagskriterien - Allgemeines .....</b>	<b>29</b>
<b>5.2</b>	<b>Zuschlagskriterien - Benchmarks .....</b>	<b>29</b>
5.2.1	Rahmenbedingungen .....	29
5.2.2	CPU Benchmarks .....	30
5.2.2.1	HPCC - Erlaubte und nicht erlaubte Änderungen .....	30
5.2.2.2	HPCC - Kompilierung .....	30
5.2.2.3	HPCC - Durchführung .....	31
5.2.2.4	HPCC - Einzureichende Werte .....	31

5.2.2.5	HPCG - Erlaubte und nicht erlaubte Änderungen .....	31
5.2.2.6	HPCG - Kompilierung .....	32
5.2.2.7	HPCG - Durchführung .....	32
5.2.2.8	HPCG - Einzureichende Werte .....	33
5.2.3	GPU Benchmarks .....	33
5.2.3.1	osu_bibw - Erlaubte und nicht erlaubte Änderungen .....	33
5.2.3.2	osu_bibw - Kompilierung .....	33
5.2.3.3	osu_bibw - Durchführung .....	33
5.2.3.4	osu_bibw - Einzureichende Werte .....	34
5.2.4	Storage Benchmarks .....	34
5.2.4.1	IO500 - Zusätzliche Rahmenbedingungen .....	35
5.2.4.2	IO500 - Erlaubte und nicht erlaubte Änderungen .....	35
5.2.4.3	IO500 - Kompilierung .....	35
5.2.4.4	IO500 - Durchführung .....	35
5.2.4.5	IO500 - Einzureichende Werte .....	36
<b>5.3</b>	<b>Zuschlagskriterien - Bewertungsmatrix .....</b>	<b>38</b>

# 1 Vorbemerkungen

Der Bieter hat zu gewährleisten, dass mit seinem Angebot sämtliche Festlegungen und Forderungen der Vergabeunterlagen erfüllt werden. Bei Zuschlag wird das Angebot Bestandteil des Vertrages werden.

Enthalten die Vergabeunterlagen nach Auffassung des Bieters Unklarheiten, die eine Preisermittlung bzw. Angebotserstellung insgesamt beeinflussen könnten, so hat der Bieter die ausschreibende Stelle darauf hinzuweisen.

Sämtliche Kommunikation im Rahmen des Vergabeverfahrens findet ausschließlich in deutscher Sprache über das Vergabetool statt. Verstöße gegen diese Kommunikationsregel (z.B. telefonische Kontaktaufnahmen) können als Verletzung vergaberechtlicher Grundsätze bewertet werden (Wettbewerbsprinzip, Gleichbehandlungs- und Transparenzgebot) und zum Ausschluss aus dem Verfahren führen.

Für das Angebot sind nur die von der Vergabestelle zur Verfügung gestellten Vergabeunterlagen zu verwenden.

Dem Bieter obliegt die Pflicht zur Vollständigkeitsprüfung der Vergabeunterlagen.

Der Bieter bestätigt mit Abgabe des Angebots, dass nach seiner fachlichen Expertise die Leistungen in der Leistungsbeschreibung abschließend und erschöpfend beschrieben und im Preisblatt vollständig aufgeführt sind und insbesondere auch keine Teilleistungen fehlen, die zur einwandfreien Erfüllung der Leistungen notwendig sind.

Nach Zuschlag durch den Auftragnehmer angesetzte Mehraufwendungen oder Zuschläge aufgrund fehlender oder fehlerhafter Vergabeunterlagen und/oder durch den Auftragnehmer nicht beschaffter Ortskenntnisse werden seitens des Auftraggebers nicht anerkannt.

Die Angebotspreise (in Euro) sind Festpreise für den Ausführungszeitraum und müssen die Kosten für Verpackung und Transport enthalten. Sofern Sie Skonto anbieten, ist dieser mit einer Zahlungsfrist von 14 Tagen zu gewähren.

Für Verpackungen gilt die zurzeit geltende Verpackungsordnung und die Verpflichtung sämtliche Verpackungen unentgeltlich zurückzunehmen.

Der zukünftige Auftragnehmer verpflichtet sich, die Lieferung am vereinbarten Lieferort gebrauchsfertig /

funktionsbereit zu übergeben.

Für das Angebot sind die von der Vergabestelle auf der eVergabe-Plattform zur Verfügung gestellten Vergabeunterlagen zu verwenden. Das Angebot ist einschließlich aller Anlagen in deutscher Sprache elektronisch über die genannte eVergabe-Plattform abzugeben.

Sofern fremdsprachige Nachweise - dazu gehören auch Datenblätter oder ähnliches – eingereicht werden, sind jeweils Übersetzungen in deutscher Sprache beizufügen. Auf ausdrückliches Verlangen der Vergabestelle hat der Bieter die Übersetzung durch einen in der Bundesrepublik Deutschland für die jeweilige Sprache amtlich vereidigten Übersetzer nachzureichen.

Eingereichte eigene Anlagen sind mit Namen zu bezeichnen, die den Inhalt wiedergeben.

Angebote können nur über die eVergabe-Plattform eingereicht werden. Die Angebotseinreichung wird über die eVergabe-Plattform bestätigt.

Angebote können nur bis zum festgelegten Ende der Angebotsfrist eingereicht werden.

Bitte beachten Sie, dass Ihr Angebot bei Nichtbestehen der formalen Angebotswertung nicht weiter gewertet werden kann.

Sie sind berechtigt, die von Ihnen bezogenen Unterlagen (insbesondere Preisangebotsvordrucke und Leistungsbeschreibungsdrucke) zwecks Abgabe Ihrer Angebote zu vervielfältigen. Sie dürfen jedoch in keinem Fall Veränderungen an den Unterlagen vornehmen. Änderungen und Ergänzungen an den Ihnen übersendeten Vergabeunterlagen sind unzulässig und führen zum Ausschluss des Angebotes von der Wertung. Dies gilt auch für die dem Angebot beigelegten Allgemeinen Geschäfts-, Liefer- oder Zahlungsbedingungen des Bieters oder eines möglicherweise eingesetzten Unterauftragnehmers.

**Eventuelle Fragen zum Vergabeverfahren sind unverzüglich nach Erhalt der Unterlagen und nur elektronisch über das Vergabeportal bis spätestens 6 Werktage vor Ablauf der Angebotsfrist zu stellen.**

Bis zum Ablauf der Angebotsfrist kann das Angebot zurückgezogen werden. Danach sind Sie bis zum Ablauf der Bindefrist an Ihr Angebot gebunden.

Unter den Angeboten, die nicht ausgeschlossen werden, wird der Zuschlag auf das Angebot mit dem besten Preis - Leistungsverhältnis erteilt. Dabei wird die erweiterte Richtwertmethode angewendet. Nur Angebote von Bietern, deren Eignung für die nachgefragte Leistung erwiesen ist, die die formellen und technischen Anforderungen erfüllen und deren Angebotspreis angemessen ist, kommen in die engere Wahl.

Ist bereits jetzt oder wird im Laufe des Vergabeverfahrens die Eröffnung des Insolvenzverfahrens oder eines vergleichbaren gesetzlichen Verfahrens über das Vermögen des Bieters eröffnet oder beantragt oder dieser Antrag mangels Masse abgelehnt oder befindet sich der Bieter bereits jetzt oder im Laufe des Vergabeverfahrens in Liquidation oder stellt er seine Tätigkeit ein, so ist dies unverzüglich mitzuteilen. Ebenso mitzuteilen ist jeder Umstand, der eine bzw. mehrere Erklärung/en des Angebotes nachträglich in Frage stellt.

## 2 Ziel der Beschaffung

Ausgeschrieben wird ein High Performance Cluster welcher den wissenschaftlichen Rechenbedarf der Martin-Luther-Universität Halle-Wittenberg in den kommenden 5 Jahren abdecken muss.

Neben der Lieferung von Hard- und Software, ist der Einbau, die Verkabelung und die Konfiguration der Hardware, die Installation und Einrichtung der angefragten Software sowie der Nachweis der Betriebsbereitschaft in Form einer Endabnahme anzubieten. Die Kriterien, wie das zu erfolgen hat, sind im Abschnitt „Leistungsbeschreibung“ beschrieben.

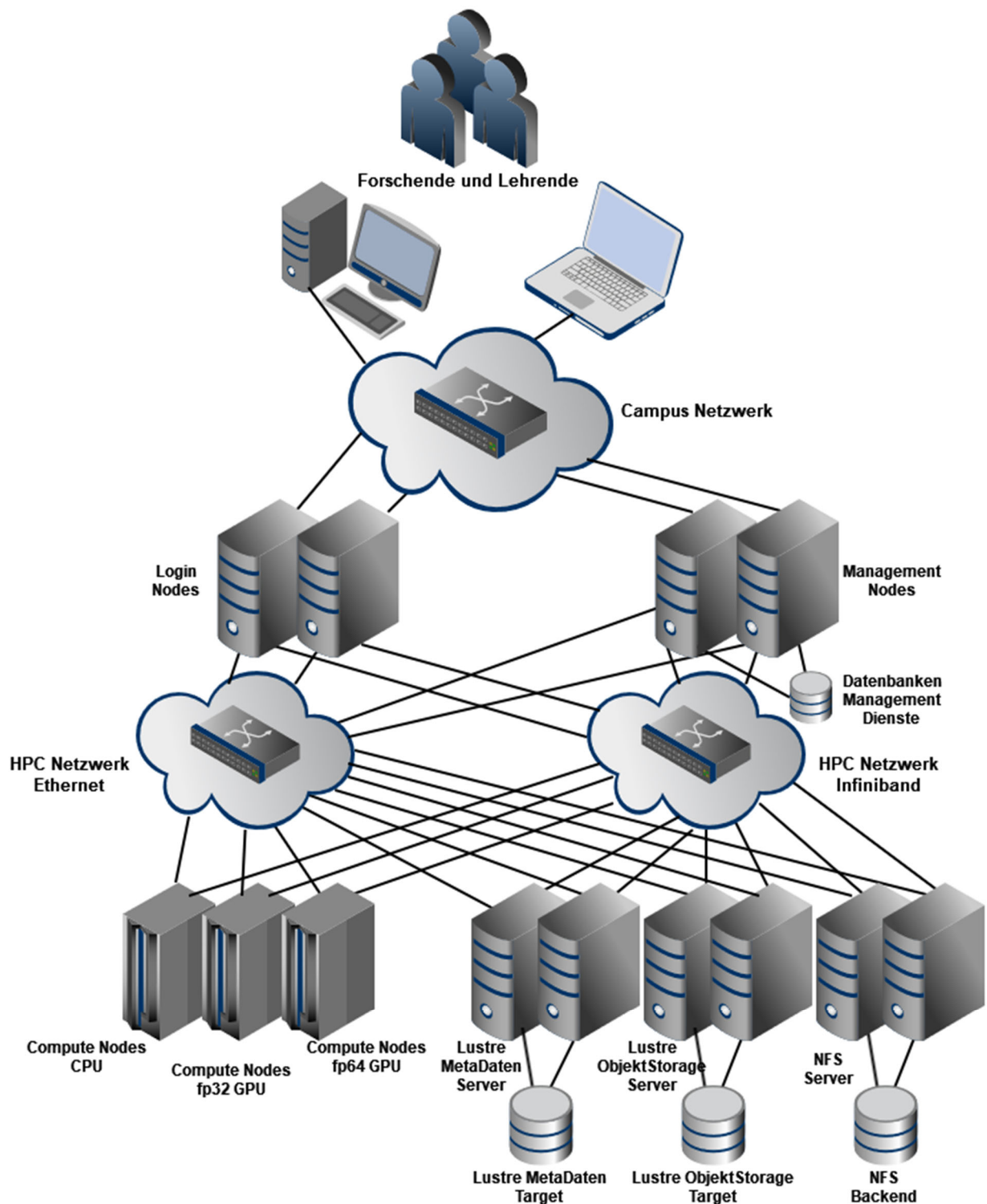
Weiterhin sind die Dokumentation aller Strukturen, Parameter und Abläufe im HPC Cluster, Beratungsleistungen sowie Schulungen anzubieten. Die Kriterien, wie das zu erfolgen hat, sind ebenfalls im Abschnitt „Leistungsbeschreibung“ beschrieben

**Das Ziel der Beschaffung ist es, das beste Preis - Leistungsverhältnis unter Einhaltung einer vorgegebenen Energieeffizienz zu erreichen.** Die Zuschlagskriterien werden in Abschnitt „Zuschlagskriterien“ erläutert.

## 3 Leistungsbeschreibung

### 3.1 Struktur des HPC Clusters

#### Struktur des HPC Clusters



Je nachdem, wofür die wissenschaftliche Software ausgelegt ist, soll diese auf CPUs, GPUs mit einfacher Fließkommagenauigkeit (32bit) oder GPUs mit doppelter Fließkommagenauigkeit (64bit) laufen. Für jede dieser drei Kategorien muss es je einen speziell ausgelegten Rechenknotentyp geben. Das sind **Compute Node CPU**, **Compute Node GPU fp32** und **Compute Node GPU fp64**.

Auf den GPU fp32 und fp64 Nodes finden **auch Berechnungen auf der CPU** statt. Deshalb sind die CPUs ebenso leistungsfähig und der RAM ebenso groß zu dimensionieren, wie bei den CPU Nodes. Weiterhin wird ein **gemeinsam genutztes Speichersystem** benötigt, das einen **hochperformanten Bereich** für die temporären Daten bei den Berechnungen (Scratch) und einen **weniger performanten Bereich** für die Zwischenspeicherung von Ausgangsdaten und berechneten Ergebnissen beinhaltet (Home).

Die HPC internen Dienste und die Datensicherung müssen auf den **Management Nodes** laufen. Der Zugang zum HPC durch die Anwender hat über die **Login Nodes** zu erfolgen.

Alle genannten Knoten sowie das Speichersystem des HPC Systems müssen sowohl durch ein **Ethernet-** als auch durch ein **Infiniband - Netzwerk** miteinander verbunden sein.

Alle angebotenen Hardwarekomponenten sind aufgebaut, verkabelt und mit installierter sowie konfigurierter Software betriebsbereit zu übergeben. Weiterhin ist eine Dokumentation anzufertigen und die betreuenden Mitarbeiter zu schulen.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Struktur des HPC Clusters“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

## **3.2 Gesamter HPC Cluster**

### **3.2.1 Kaufmännische Kriterien**

Alle in diesem Abschnitt genannten kaufmännischen Kriterien sind einzuhalten. Bei Nichterfüllung dieser Kriterien, wird das Angebot ausgeschlossen.

Der angebotene Gesamtpreis inkl. MwSt. muss unter **3.000.000 Euro liegen**. Bei Überschreiten dieses Maximalbudgets kann kein Zuschlag erteilt werden.

Der Bieter muss zum Zeitpunkt der Ausschreibung mind. **drei vergleichbare Referenzinstallationen in den letzten 5 Jahren** mit einem Gesamtpreis inkl. MwSt. von jeweils **mind. 2.500.000 Euro** vorweisen. Das soll gewährleisten, dass der Bieter über die notwendigen Mitarbeiter in Anzahl und Qualifikation sowie über das notwendige Wissen zum Aufbau und Inbetriebnahme eines HPC Systems in der ausgeschriebenen Größenordnung verfügt.

Der Bieter muss zum Zeitpunkt der Ausschreibung mind. eine **DIN ISO9001:2015 Zertifizierung** vorweisen. Das soll gewährleisten, dass der Bieter über die zum Aufbau und Inbetriebnahme eines HPC Systems notwendigen organisatorischen Strukturen verfügt sowie in der Lage ist, die geforderte Garantie angemessen zu erfüllen.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Kaufmännische Kriterien“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### **3.2.2 Technische Kriterien**

Alle in diesem Abschnitt genannten technischen Kriterien sind einzuhalten. Bei Nichterfüllung dieser Kriterien, wird das Angebot ausgeschlossen.

Es sind **ausschließlich** Neugeräte zu liefern. Wiederaufbereitete Produkte, Rückläufer oder Graumimporte sind ausgeschlossen.

Die maximale Leistungsaufnahme des gesamten Clusters muss unterhalb **150 kW liegen**. Das entspricht der maximal verfügbaren Kälteleistung am Ort der Aufstellung. Bei Überschreiten dieser Grenze kann kein Zuschlag erteilt werden.

Für sämtliche gelieferte **Hardware** muss für mindestens **5 Jahre ab dem Zeitpunkt der Endabnahme** eine **Garantie inklusive vor - Ort Service** geleistet werden.

Das bedeutet, dass der Auftragnehmer nach Meldung einer Hardwarestörung **deren Ursache zu ermitteln hat und alle notwendigen Schritte bis zur Wiederherstellung der Funktion des defekten Bauteils ergreifen muss**. Das beinhaltet gegebenenfalls den Austausch des Bauteils und muss, wenn nötig, auch vor Ort erfolgen.

Die Kommunikation mit dem Garantiedienstleister muss über eine **deutschsprachige Telefon - Hotline** bzw. **deutschsprachigen E-Mail Austausch** erfolgen.

Die **Reaktionszeit** auf eine Störungsmeldung darf höchstens **einen Arbeitstag von Montag bis Freitag** (außer an gesetzlichen Feiertagen) betragen. An jedem dieser Tage muss der Garantiedienstleister **mindestens 8 Stunden** lang erreichbar sein.

Kriterium	Spezifikation	Erläuterung
Vorhandene Racks	7 Stück Rack Rittal VX IT 5309.816 Tiefe 100 cm  Keine Zwischenwände zwischen den Racks vorhanden - eine direkte Verkabelung ist möglich  Jedes Rack ist mit je 2 PDU Rittal DK 7979.537 ausgestattet.	Ein Überstehen über das hintere Rackende und das Weglassen der Hintertür ist erlaubt
Maximal belegbare Höheneinheiten	294	entspricht 7 Racks Rittal VX IT 5309.816 zu je 42 HE
Formfaktor jedes Gehäuses von Nodes / Storage / Netzwerk	Rackmount Gehäuse / Einbaufähig in 19" Rack	Desktop – Technik ist nicht zugelassen
Maximale Leistungsaufnahme für den gesamten HPC Cluster	150 kW	entspricht der geplanten Kühlleistung für die 7 Racks Luftkühlung in eingehaustem Kaltgang
Maximaler Leistungsaufnahme je Rack <b>ohne</b> redundante Stromversorgung aller Komponenten	40 kVA	entspricht 2 PDU Rittal DK 7979.537 mit je 3 Phasen mit je 32A / 230V = ca 40 kVA <b>gilt für die Compute Nodes</b>
Maximale Leistungsaufnahme je Rack <b>mit</b> redundanter Stromversorgung aller Komponenten	20 kVA	entspricht 1 PDU Rittal DK 7979.537 mit je 3 Phasen mit je 32A / 230V = ca. 20 kVA <b>gilt für die zentralen Nodes, den Storage, das Netzwerk</b>
Compute Nodes	Keine redundante Auslegung notwendig  Stateless, d.h. Boot des Betriebssystems über HPC Ethernet Netzwerk	
Hersteller der Compute Nodes	ein einheitlicher Hersteller	Ermöglicht eine einheitliche Betriebssysteminstallation inkl.



		Managementsoftware und Treiber
CPUs der Compute Nodes	ein einheitliches CPU Modell, Serverprozessor, x64 kompatibel	ein einheitlicher CPU Befehlssatz mit alle Features. Dadurch ist nur ein „Kompilat“ pro HPC Software nötig.
PCIe Karten der Compute-, Management- und Login- Nodes	Jede PCIe Steckkarte muss mit mind. PCIe Gen 4 betrieben werden.	Gewährleistet eine hohe IO Leistung
Management und Login Nodes sowie Storage Server	<b>innerhalb eines Nodes</b> Redundanz bei RAM (ECC), Datenträgern (Raid), Ethernet - Anbindung (Failover) sowie Stromversorgung. <b>Systemweit</b> Kalte Redundanz <sup>1</sup> der Management-, Login und Storage- Nodes	gewährleistet erhöhte Ausfallsicherheit
HPC Netzwerk Infiniband und Ethernet	Redundanz der Stromversorgung	gewährleistet erhöhte Ausfallsicherheit
Verkabelung HPC Netzwerk Infiniband und Ethernet	Zu jeder benötigten Netzwerkanbindung ist ein in Art und Beschaffenheit passendes Kabel mitzuliefern.	

### Definition des Begriffes „kalte Redundanz“

Bei der allgemein gebräuchlichen Verwendung des Begriffes „Redundanz“ ist die aktive bzw. heiße Redundanz der jeweiligen Funktion gemeint, d.h. fällt eine der redundanten Komponenten aus, erfolgt keine Unterbrechung der jeweiligen Funktion.<sup>2</sup>

Mit „kalter Redundanz“ ist hier gemeint, dass der Ausfall einer redundanten Komponente zunächst zu einem Ausfall der zugehörigen Funktion führt, diese aber nach maximal 5 Minuten durch eine Failover – Software wieder zur Verfügung gestellt wird.

### Definition des Begriffes „effektiv nutzbare Netto Kapazität“

Mit „effektiv nutzbarer Netto Kapazität“ ist die Kapazität gemeint, die man in einem Dateisystem auf dem jeweiligen Datenträger/RAID Verbund tatsächlich lesen und beschreiben kann.

Bei den im Folgenden angegebenen Werten wird davon ausgegangen, dass die „effektiv nutzbare Netto Kapazität“ mindestens 90% der Herstellerangabe an Kapazität entspricht.

### Definition des Begriffes „active / active Betrieb“ von Storage Servern

Mit „active / active Betrieb“ von Storage Servern ist gemeint, dass alle beteiligten Storage Server gleichzeitig aktiv den jeweiligen Storagedienst anbieten. Das bedeutet, dass ein Storageclient (bzw. hier ein Compute Node) diverse Daten an jeden dieser Server senden bzw. von diesen empfangen kann. Fällt einer der Storage Server aus, übernimmt ein anderer Server dessen Dienste. Abweichend von der allgemein gebräuchlichen Definition<sup>3</sup> ist eine Ausfallzeit von max. 5 Minuten erlaubt (siehe kalte Redundanz).

<sup>1</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>2</sup> Siehe auch [https://de.wikipedia.org/wiki/Redundanz\\_\(Technik\)](https://de.wikipedia.org/wiki/Redundanz_(Technik))

<sup>3</sup> Siehe auch <https://de.wikipedia.org/wiki/Aktiv/Aktiv-Cluster>

**Definition des Begriffes „active / passive Betrieb“ von Storage Servern**

Mit „active / passive Betrieb“ von Storage Servern ist gemeint, dass nur einer der beteiligten Storage Server aktiv den Storagedienst anbietet und alle anderen sich in passiver Übernahmebereitschaft befinden<sup>4</sup>. Das bedeutet, dass ein Storageclient (bzw. hier ein Compute Node) nur an den aktiven Server Daten senden bzw. von diesem empfangen kann. Fällt der aktive Storage Server aus, übernimmt ein passiver Server dessen Dienste. Dabei kommt es zu Ausfallzeiten von max. 5 Minuten (siehe kalte Redundanz).

**Definition des Begriffes „Energieeffizienzklasse“**

Mit Energieeffizienz ist das Verhältnis von Energieertrag (Output) zur zugeführten Energie (Input) gemeint<sup>5</sup>. Die Energieeffizienz wird in dieser Ausschreibung über den elektrischen Wirkungsgrad der verwendeten Netzteile anhand der „80 Plus“ Norm für 230 V Spannung bewertet<sup>6</sup>. Die Zertifizierung innerhalb der 80 Plus Norm (Stufen 80+ Standard bis 80+ Titanium) wird ab hier als Energieeffizienzklasse bezeichnet.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Technische Kriterien“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

**3.3 Rechenknoten / Compute Nodes****3.3.1 CPU Compute Nodes**

Alle Angaben in der Spalte Komponente beziehen sich **jeweils auf einen Node**. Alle folgenden Kriterien sind für CPU Compute Nodes einzuhalten:

Komponente	Spezifikation	Erläuterung
CPU	mind. 2 Stück, Serverprozessor, x64 kompatibel  pro CPU: mind. 96 Rechenkerne, mind. 4MB L3-Cache pro Rechenkern, max. 400W TDP  Gleiches CPU Modell für alle Compute Nodes	zwei CPU zur Freischaltung von genügend RAM Steckplätzen  hohe TFlop Leistung pro CPU, um die nötige Gesamtpformance zu erreichen (siehe Bewertungsmatrix) sowie Begrenzung der abgegebenen Wärmemenge
RAM	mind. 1536GB, mit ECC  mind. DDR5, mind. 64GB Module  Betrieb der Module mit mind. 6000MT/s	um große Datenmodelle im RAM halten zu können  ECC wegen Ausfallsicherheit  64GB Module wegen Erweiterbarkeit und späterer Tauschmöglichkeit mit Management/Login Nodes
SSD	mind. 1 Stück mit mind. je 3,8 TB, NVMe, mind. 1 DWPD, GPUDirect Storage kompatibel <sup>7</sup>	als temporäres lokales Dateisystem für I/O intensive Jobs

<sup>4</sup> Siehe auch <https://de.wikipedia.org/wiki/Failover#Failover-Cluster>

<sup>5</sup> Siehe auch <https://de.wikipedia.org/wiki/Energieeffizienz>

<sup>6</sup> Siehe auch [https://de.wikipedia.org/wiki/80\\_PLUS](https://de.wikipedia.org/wiki/80_PLUS)

<sup>7</sup> Siehe <https://docs.nvidia.com/gpudirect-storage/troubleshooting-guide/index.html#det-nvme-support-gds>

	<p>mind. 150 kIOPS beim Benchmark io500.posix.mdtest-hardwrite.ssd<sup>8</sup></p> <p>Die Übertragungsleistung der SSD fließt in die Bewertung des Angebotes ein<sup>9</sup>.</p> <p>Gleiches SSD Modell für alle Compute Nodes, Erweiterbarkeit um mind. eine weitere NVMe SSD</p>	Hohe lokale Storage Performance um die maximale Gesamtleistung des Compute Nodes zu gewährleisten.
Netzwerkschnittstellen Ethernet	<p>mind. 1Port</p> <p>mind. 10 GbE</p>	zur Anbindung des Betriebssystems an das HPC Netzwerk Ethernet
Netzwerkschnittstellen Infiniband	<p>mind. 1 Port, Betrieb mit mind. PCIe Gen 4</p> <p>mind. NDR 200Gbit/s Anbindung</p> <p>inkl. Infiniband Twin-Port Splitterkabel, NDR 800Gbit/s to 4x200Gbit/s, OSFP to 4xOSFP (1 Kabel für 4 Knoten)</p>	<p>zur Anbindung mit einem Port zu 200Gbit/s an das HPC Netzwerk Infiniband</p> <p>1 Splitterkabel für 4 Nodes, d.h. rein rechnerisch „¼ Kabel pro Node“</p>
Server Management Controller (BMC)	<p>mind. 1 Ethernetport incl. Kabel, mind. 1 Gbit</p> <p>Redfish oder IPMI API</p> <p>inkl. Freischaltung der erweiterten BMC Funktionen (siehe rechts)</p>	<p>zur Anbindung des BMC an das HPC Netzwerk Ethernet</p> <p>Erweiterte BMC Funktionen sind u.a. „Serverkonsole im Browser bedienen“, „ISO Images als USB Gerät am Server mounten“ „Begrenzung des Stromverbrauchs / Powercapping“</p>
Stromversorgung	<p>mind. 1 Netzteil</p> <p>mind. Energieeffizienzklasse 80+ Titanium</p>	Redundanz der Stromversorgung erwünscht, aber nicht nötig
Sonstiges	<p>inkl. 24h Burn-In Test</p> <p>Garantie entsprechend den Festlegungen in Abschnitt „Technische Kriterien“.</p>	

**Es sind exakt 16 CPU Compute Nodes entsprechend der obigen Spezifikation anzubieten.**

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „CPU Compute Nodes“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

<sup>8</sup> muss bei der Abnahme entsprechend den Vorgaben aus Abschnitt „Zuschlagskriterien - Benchmarks“ nachgewiesen werden

<sup>9</sup> Zu Details siehe Abschnitt „Zuschlagskriterien“

### 3.3.2 GPU fp32 Compute Nodes

Alle Angaben in der Spalte Komponente beziehen sich **jeweils auf einen Node**. Alle folgenden Kriterien sind für GPU fp32 Compute Nodes einzuhalten:

Komponente	Spezifikation	Erläuterung
CPU	mind. 2 Stück, Serverprozessor, x64 kompatibel  pro CPU: mind. 96 Rechenkerne, mind. 4MB L3-Cache pro Rechenkern, max. 400W TDP  Gleiches CPU Modell für alle Compute Nodes	zwei CPU zur Freischaltung von genügend RAM Steckplätzen  hohe TFlop Leistung pro CPU, um die nötige Gesamtperformance zu erreichen (siehe Bewertungsmatrix) sowie Begrenzung der abgegebenen Wärmemenge
RAM	mind. 1536GB, mit ECC  mind. DDR5, mind. 64GB Module  Betrieb der Module mit mind. 6000MT/s	um große Datenmodelle im RAM halten zu können  ECC wegen Ausfallsicherheit  64GB Module wegen Erweiterbarkeit und späterer Tauschmöglichkeit mit Management/Login Nodes
GPU	4 GPU je Server  pro GPU: mind. 96GB RAM mit ECC, passiv gekühlt, Betrieb mit mind. PCIe 5.0 x16  CUDA und GPUDirect kompatibel	Die vorhandene wiss. Software unterstützt GPU Berechnungen nur über die CUDA Schnittstelle
SSD	mind. 1 Stück mit mind. 3,8 TB, NVMe, mind. 1 DWPD, GPUDirect Storage kompatibel <sup>10</sup>  mind. 150 kIOPS beim Benchmark io500.posix.mdtest-hardwrite.ssd <sup>11</sup>  Die Übertragungsleistung der SSD fließt in die Bewertung des Angebotes ein <sup>12</sup> .  Gleiches SSD Modell für alle Compute Nodes, Erweiterbarkeit um mind. eine weitere NVMe SSD	als temporäres lokales Dateisystem für I/O intensive Jobs  Hohe lokale Storage Performance um eine hohe Gesamtleistung des Compute Nodes zu gewährleisten.
Netzwerkschnittstellen Ethernet	mind. 1Port  mind. 10 GbE	zur Anbindung des Betriebssystems an das HPC Netzwerk Ethernet

<sup>10</sup> Siehe <https://docs.nvidia.com/gpudirect-storage/troubleshooting-guide/index.html#det-nvme-support-gds>

<sup>11</sup> muss bei der Abnahme entsprechend den Vorgaben aus Abschnitt „Zuschlagskriterien - Benchmarks“ nachgewiesen werden

<sup>12</sup> Zu Details siehe Abschnitt „Zuschlagskriterien“

Netzwerkschnittstellen Infiniband	mind. 1 Port, Betrieb mit mind. PCIe Gen 4  mind. NDR 200Gbit/s Anbindung  inkl. Infiniband Twin-Port Splitterkabel, NDR 800Gbit/s to 4x200Gbit/s, OSFP to 4xOSFP (1 Kabel für 4 Knoten)	zur Anbindung mit einem Port zu 200Gbit/s an das HPC Netzwerk Infiniband  1 Splitterkabel für 4 Nodes, d.h. rein rechnerisch „1/4 Kabel pro Node“
Server Management Controller (BMC)	mind. 1 Ethernetport incl. Kabel, mind. 1 Gbit  Redfish oder IPMI API  inkl. Freischaltung der erweiterten BMC Funktionen (siehe rechts)	zur Anbindung des BMC an das HPC Netzwerk Ethernet  Erweiterte BMC Funktionen sind u.a. „Serverkonsole im Browser bedienen“, „ISO Images als USB Gerät am Server mounten“ „Begrenzung des Stromverbrauchs / Powercapping“
Stromversorgung	mind. 1 Netzteil  mind. Energieeffizienzklasse 80+ Titanium	Redundanz der Stromversorgung erwünscht, aber nicht nötig
Sonstiges	inkl. 24h Burn-In Test  Garantie entsprechend den Festlegungen in Abschnitt „Technische Kriterien“.	

Die Anzahl der GPU fp32 Compute Nodes fließt in die Bewertung des Angebotes ein.<sup>13</sup>

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „GPU fp32 Compute Nodes“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.3.3 GPU fp64 Compute Nodes

Alle Angaben in der Spalte Komponente beziehen sich **jeweils auf einen Node**. Alle folgenden Kriterien sind für GPU fp64 Compute Nodes einzuhalten:

Komponente	Spezifikation	Erläuterung
CPU	mind. 2 Stück, Serverprozessor, x64 kompatibel  pro CPU: mind. 96 Rechenkerne, mind. 4MB L3-Cache pro Rechenkern, max. 400W TDP  Gleiches CPU Modell für alle Compute Nodes	zwei CPU zur Freischaltung von genügend RAM Steckplätzen  hohe TFlop Leistung pro CPU, um die nötige Gesamtleistung zu erreichen (siehe Bewertungsmatrix) sowie Begrenzung der abgegebenen Wärmemenge
RAM	mind. 1536GB, mit ECC	um große Datenmodelle im RAM halten zu können

<sup>13</sup> Zu Details siehe im Anhang „Bewertungsmatrix“

	mind. DDR5, mind. 64GB Module  Betrieb der Module mit mind. 6000MT/s	ECC wegen Ausfallsicherheit  64GB Module wegen Erweiterbarkeit und späterer Tauschmöglichkeit mit Management/Login Nodes
GPU	4 GPU je Server  pro GPU: mind. 141GB RAM mit ECC, passiv gekühlt, CUDA und GPUDirect kompatibel  einsatzbereiter NVLink für alle 4 GPUs (4-way NVLink)	Die vorhandene wiss. Software unterstützt GPU Berechnungen nur über die CUDA Schnittstelle  „einsatzbereiter NVLink“ bedeutet: inkl. Zusatzhardware
SSD	mind. 1 Stück mit mind. 3,8 TB, NVMe, mind. 1 DWPD, GPUDirect Storage kompatibel <sup>14</sup>  mind. 150 kIOPS beim Benchmark io500.posix.mdtest-hardwrite.ssd <sup>15</sup>  Die Übertragungsleistung der SSD fließt in die Bewertung des Angebotes ein <sup>16</sup> .  Gleiches SSD Modell für alle Compute Nodes, Erweiterbarkeit um mind. eine weitere NVMe SSD	als temporäres lokales Dateisystem für I/O intensive Jobs  Hohe lokale Storage Performance um eine hohe Gesamtleistung des Compute Nodes zu gewährleisten.
Netzwerkschnittstellen Ethernet	mind. 1Port  mind. 10 GbE	zur Anbindung des Betriebssystems an das HPC Netzwerk Ethernet
Netzwerkschnittstellen Infiniband	mind. 1 Port, Betrieb mit mind. PCIe Gen 4  mind. NDR 200Gbit/s Anbindung  inkl. Infiniband Twin-Port Splitterkabel, NDR 800Gbit/s to 4x200Gbit/s, OSFP to 4xOSFP (1 Kabel für 4 Knoten)	zur Anbindung mit einem Port zu 200Gbit/s an das HPC Netzwerk Infiniband  1 Splitterkabel für 4 Nodes, d.h. rein rechnerisch „1/4 Kabel pro Node“
Server Management Controller (BMC)	mind. 1 Ethernetport incl. Kabel, mind. 1 Gbit  Redfish oder IPMI API  inkl. Freischaltung der erweiterten BMC Funktionen (siehe rechts)	zur Anbindung des BMC an das HPC Netzwerk Ethernet  Erweiterte BMC Funktionen sind u.a. „Serverkonsole im Browser bedienen“, „ISO Images als USB Gerät am Server mounten“

<sup>14</sup> Siehe <https://docs.nvidia.com/gpudirect-storage/troubleshooting-guide/index.html#det-nvme-support-gds>

<sup>15</sup> muss bei der Abnahme entsprechend den Vorgaben aus Abschnitt „Zuschlagskriterien - Benchmarks“ nachgewiesen werden

<sup>16</sup> Zu Details siehe Abschnitt „Zuschlagskriterien“

		„Begrenzung des Stromverbrauchs / Powercapping“
Stromversorgung	mind. 1 Netzteil mind. Energieeffizienzklasse 80+ Titanium	Redundanz der Stromversorgung erwünscht, aber nicht nötig
Sonstiges	inkl. 24h Burn-In Test Garantie entsprechend den Festlegungen in Abschnitt „Technische Kriterien“.	

Es sind exakt 8 GPU fp64 Compute Nodes entsprechend der obigen Spezifikation anzubieten.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „GPU fp64 Compute Nodes“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.4 Zentrale Nodes

#### 3.4.1 Management Nodes

Alle Angaben in der Spalte Komponente beziehen sich **jeweils auf einen Node**. Alle folgenden Kriterien sind für Management Nodes einzuhalten:

Komponente	Spezifikation	Erläuterung
CPU	mind. 2 Stück, Serverprozessor, x64 kompatibel pro CPU: mind. 24 Rechenkerne, mind. 3,5 GHz Basisakt/base frequency	zwei CPU zur Freischaltung von RAM Steckplätzen Hohe single thread Performance für die Ausführung der HPC Management Software und der Datensicherung
RAM	mind. 768GB, mit ECC mind. DDR5, mind. 32GB Module Betrieb der Module mit mind. 6000MT/s	ECC wegen Ausfallsicherheit
SSD lokal	mind. 2 Stück mit mind. 1,9 TB, NVMe, mind. 1 DWPD betrieben an Hardware RAID Controller	RAID1 für das <b>Betriebssystem</b>
Storage Controller	mind. 1 SAS-4/SAS 24G HBA, mind. PCIe 4.0 x8  <b>ACHTUNG: Bei Kombination aus Management Node und NFS Server muss ggf. ein</b>	zur Anbindung von mind. einem geteilten Festplattensystem für die <b>HPC Managementdaten</b>  zur Anbindung von mind. einem geteilten Festplattensystem für den <b>Home/NFS Bereich</b>

	<b>zusätzlicher SAS-4/SAS 24G HBA eingebaut werden.</b>	<b>Für bessere Performance erwünscht: SAS HBA PCIe 4.0 x16</b>
Netzwerkschnittstellen Ethernet	mind. 2 Ports SFP28 mit jeweils mind. 25GbE Anbindung pro Port: mind. ein SFP28 multimode/shortwave LWL Transceiver für Server NIC und mind. ein SFP28 „Cisco Nexus compatible“ multimode/shortwave LWL Transceiver für Switch  mind. 2 Ports mit mind. je 10GbE	zur redundanten Anbindung an das Cisco Nexus Campus Netzwerk der Auftraggeberin mit 25 GbE  zur redundanten Anbindung an das HPC Netzwerk Ethernet
Netzwerkschnittstellen Infiniband	mind. 1 Port, Betrieb mit mind. PCIe Gen 4  mind. NDR 200Gbit/s Anbindung  inkl. Infiniband Twin-Port Splitterkabel, NDR 800Gbit/s to 4x200Gbit/s, OSFP to 4xOSFP (1 Kabel für 4 Knoten)	zur Anbindung mit einem Port zu 200Gbit/s an das HPC Netzwerk Infiniband  1 Splitterkabel für 4 Nodes, d.h. rein rechnerisch „¼ Kabel pro Node“
Server Management Controller (BMC)	mind. 1 Ethernetport incl. Kabel, mind. 1 Gbit  Redfish oder IPMI API  inkl. Freischaltung der erweiterten BMC Funktionen (siehe rechts)	zur Anbindung des BMC an das HPC Netzwerk Ethernet  Erweiterte BMC Funktionen sind u.a. „Serverkonsole im Browser bedienen“, „ISO Images als USB Gerät am Server mounten“ „Begrenzung des Stromverbrauchs / Powercapping“
Stromversorgung	mind. 2 hotswap Netzteile  mind. Energieeffizienzklasse 80+ Titanium	Redundante Stromversorgung
Sonstiges	inkl. 24h Burn-In Test  Garantie entsprechend den Festlegungen in Abschnitt „Technische Kriterien“.	

**Es sind mindestens zwei Management Nodes anzubieten.**

Es ist **mindestens ein shared Festplattensystem/JBOD mit redundanten Controllern für die HPC Managementdaten** anzubieten. Weiteres zu dessen Beschaffenheit in Abschnitt „Managementdaten Storage“.

Wenn die Management Nodes gleichzeitig auch NFS Server sind, ist **mindestens ein weiteres shared Festplattensystem/JBOD mit redundanten Controllern für das Backend des NFS Servers bzw. die Daten des Home Bereiches** anzubieten. Weiteres zu dessen Beschaffenheit in Abschnitt „NFS Server und NFS Backend Storage“.



- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Management Nodes“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.4.2 Login Nodes

Alle Angaben in der Spalte Komponente beziehen sich **jeweils auf einen Node**. Alle folgenden Kriterien sind für Login Nodes einzuhalten:

Komponente	Spezifikation	Erläuterung
CPU	mind. 2 Stück, Serverprozessor, x64 kompatibel  pro CPU: mind. 24 Rechenkerne, mind. 3,5 GHz Basisakt/base frequency	zwei CPU zur Freischaltung von RAM Steckplätzen Hohe single thread Performance für das Kompilieren und Testen von Software
RAM	mind. 768GB, mit ECC  mind. DDR5, mind. 32GB Module  Betrieb der Module mit mind. 6000MT/s	ECC wegen Ausfallsicherheit
SSD	mind. 2 Stück mit mind. 1,9 TB, NVMe, mind. 1 DWPD  betrieben an Hardware RAID Controller	RAID1 für das <b>Betriebssystem</b>
Netzwerkschnittstellen Ethernet	mind. 2 Ports SFP28 mit jeweils mind. 25GbE Anbindung pro Port: mind. ein SFP28 multimode/shortwave LWL Transceiver für Server NIC und mind. ein SFP28 „Cisco Nexus compatible“ multimode/shortwave LWL Transceiver für Switch  mind. 2 Ports mit mind. je 10GbE	zur redundanten Anbindung an das Cisco Nexus Campus Netzwerk der Auftraggeberin mit 25 GbE  zur redundanten Anbindung an das HPC Netzwerk Ethernet
Netzwerkschnittstellen Infiniband	mind. 1 Port, Betrieb mit mind. PCIe Gen 4  mind. NDR 200Gbit/s Anbindung  inkl. Infiniband Twin-Port Splitterkabel, NDR 800Gbit/s to 4x200Gbit/s, OSFP to 4xOSFP (1 Kabel für 4 Knoten)	zur Anbindung mit einem Port zu 200Gbit/s an das HPC Netzwerk Infiniband  1 Splitterkabel für 4 Nodes, d.h. rein rechnerisch „¼ Kabel pro Node“
Server Management Controller (BMC)	mind. 1 Ethernetport incl. Kabel, mind. 1 Gbit	zur Anbindung des BMC an das HPC Netzwerk Ethernet

	Redfish oder IPMI API inkl. Freischaltung der erweiterten BMC Funktionen (siehe rechts)	Erweiterte BMC Funktionen sind u.a. „Serverkonsole im Browser bedienen“, „ISO Images als USB Gerät am Server mounten“ „Begrenzung des Stromverbrauchs / Powercapping“
Stromversorgung	mind. 2 hotswap Netzteile mind. Energieeffizienzklasse 80+ Titanium	Redundante Stromversorgung
Sonstiges	inkl. 24h Burn-In Test  Garantie entsprechend den Festlegungen in Abschnitt „Technische Kriterien“.	

Es sind mindestens zwei Login Nodes anzubieten.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Login Nodes“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.5 Storage

#### 3.5.1 Storage - Allgemein

- Das von den Compute-, Management- und Login - Nodes gemeinsam genutzte Storage System muss in einen hochperformanten **Bereich Scratch** für die temporären Daten der Berechnungen auf den Compute Nodes und einen weniger performanten **Bereich Home** für die Zwischenspeicherung von Ausgangsdaten und berechneten Ergebnissen unterteilt sein.
- Die Anbindung dieser **beiden Storagebereiche** an die Nodes muss **über das HPC Infiniband Netzwerk** erfolgen.
- Beide Storagebereiche müssen **kalt redundant**<sup>17</sup> ausgelegt sein. Das heißt, wenn ein Storage - Server ausfällt, fällt zunächst auch der jeweilige Storage - Bereich (Home oder Scratch) aus. Dieser muss aber nach maximal 5 Minuten wieder zur Verfügung stehen.
- Der Home Bereich darf **nur für die Login- und Management Nodes beschreibbar** sein. **Für die Compute Nodes** wird der Home Bereich **nur lesend** freigegeben. Das soll verhindern, dass Jobs aktiv Daten auf dem Home Bereich ablegen und diesen überlasten. Die Übertragung der Ergebnisdaten von Scratch nach Home hat durch **Slurm Burst Buffer Staging**<sup>18</sup> zu erfolgen.
- Eine Datensicherung erfolgt **nur für den Home Bereich**. Das hat über die Management Nodes zu erfolgen und muss mithilfe von Dateisystem Snapshots möglich sein. Die Einrichtung der Datensicherung übernimmt die Auftraggeberin.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Storage - Allgemein“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

#### 3.5.2 Storage Software

<sup>17</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>18</sup> Siehe [https://slurm.schedmd.com/burst\\_buffer.html](https://slurm.schedmd.com/burst_buffer.html)

- Für den Scratch Bereich muss die Open Source Software „**Lustre**“ verwendet werden. Diese Forderung besteht, weil Lustre auch nach Ablauf der Software Subscription mit allen Funktionen weiterbetrieben werden darf.  
(Im Gegensatz dazu müssen zum Beispiel die „BeeGFS Enterprise Features“ nach Ablauf der Subscription deaktiviert werden.)
- Alle eingesetzten Lustre MetaDatenServer und ObjectStorageServer müssen für den **active / active Betrieb**<sup>19</sup> konfiguriert werden.
- Für den Home Bereich muss **NFS** zum Einsatz kommen. Das entkoppelt den Home vom Scratch Bereich und sorgt so für mehr Stabilität. Weiterhin ist auf diese Weise eine Datensicherung über Filesystem Snapshots möglich.
- Die NFS Server müssen für den **active / passive Betrieb**<sup>20</sup> konfiguriert werden.
- Die zusätzliche Gewährung einer **Garantie auf die Lustre Software über 5 Jahre** fließt in die Bewertung des Angebotes ein.<sup>21</sup>

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Storage Software“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.5.3 Storage Hardware

#### 3.5.3.1 Storage Hardware - Allgemein

- Ein **Mischbetrieb von HDDs und SSDs** im selben Plattensystem ist nicht zulässig.
- Der **Ausfall jeder einzelnen SSD bzw. HDD** im Storage Bereich muss durch **RAID** abgesichert sein (Details siehe in den folgenden Abschnitten).
- Bei allen geteilten Festplattensystemen müssen **dual-port SSDs bzw. HDDs** verwendet werden. Weiterhin müssen diese im laufenden Betrieb austauschbar (hot-swap) sein.
- Jedes Festplattensystem muss intern mind. **zwei redundante Controller und zwei redundante Netzteile** besitzen.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Storage Hardware - Allgemein“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

#### 3.5.3.2 Lustre MetaDatenServer und ObjectStorageServer

- Es ist zulässig, einen Lustre MetaDatenServer und einen ObjektStorageServer auf **derselben Serverhardware** zu betreiben, wenn im Fehlerfall **mindestens ein weiterer kombinierter Meta-Daten/ObjektStorage Server** zur Verfügung steht, der dessen Funktion übernimmt.
- Für jeden Lustre MetaDatenServer bzw. ObjektStorageServer muss eine kalte Redundanz<sup>22</sup> gewährleistet sein.

Alle Angaben in der Spalte Komponente beziehen sich **jeweils auf einen Server/Node/Controller**. Alle folgenden Kriterien sind für Lustre MetaDatenServer sowie ObjektStorageServer einzuhalten:

Komponente	Spezifikation	Erläuterung
RAM	mind. 256GB, mind. DDR4, mit ECC	ECC wegen Ausfallsicherheit

<sup>19</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>20</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>21</sup> Zu Details siehe im Anhang „Bewertungsmatrix“

<sup>22</sup> Definition siehe Abschnitt „Technische Kriterien“

Controller	<p><b>Wenn der Server MetaDaten-Server ist:</b> mind. ein herstellerspezifischer PCIe Bus zum Sharen von NVMe SSDs zwischen zwei Servern/Nodes/Controllern<sup>23</sup> <b>bzw.:</b> mind. 1 SAS-4/SAS 24G HBA, mind. PCIe 4.0 x8</p> <p><b>Wenn der Server ObjektStorageServer ist:</b> mind. 2 Controller SAS-4/SAS 24G, mind. PCIe 4.0 x8</p>	<p>PCIe Bus für sehr hohe IO Performance zur Anbindung von mind. einem geteilten Festplattensystem als <b>MDT</b> Für bessere Performance erwünscht: SAS HBA PCIe 4.0 <b>x16</b></p> <p>zur Anbindung von mind. zwei geteilten Festplattensystemen als <b>OST</b>, pro Festplattensystem ein separater SAS HBA</p>
Netzwerkschnittstellen Ethernet	mind. zwei Ports mit mind. je 10GbE	zur redundanten Anbindung an das HPC Netzwerk Ethernet
Netzwerkschnittstellen Infiniband	<p>mind. ein Port, Betrieb mit mind. PCIe Gen 4</p> <p>mind. NDR 200Gbit/s Anbindung</p> <p>inkl. Infiniband Twin-Port Splitterkabel, NDR 800Gbit/s to 4x200Gbit/s, OSFP to 4xOSFP (1 Kabel für 4 Knoten) bzw. NDR 800Gbit/s to 2x400Gbit/s, OSFP to 2xOSFP (1 Kabel für 2 Knoten)</p>	<p>zur Anbindung mit einem Port zu 200Gbit/s an das HPC Netzwerk Infiniband Für bessere Performance erwünscht: <b>zwei</b> Ports zu je NDR 200Gbit/s bzw. ein Port zu <b>NDR 400Gbit/s</b></p> <p>1 Splitterkabel für 4 Nodes, d.h. rein rechnerisch „¼ Kabel pro Node“</p>
Server Management Controller (BMC)	<p>mind. 1 Ethernetport incl. Kabel, mind. 1 Gbit</p> <p>Redfish oder IPMI API</p>	zur Anbindung des BMC an das HPC Netzwerk Ethernet
Stromversorgung	<p>mind. 2 hotswap Netzteile</p> <p>mind. Energieeffizienzklasse 80+ Platinum</p>	Redundante Stromversorgung
Sonstiges	<p>inkl. 24h Burn-In Test</p> <p>Garantie entsprechend den Festlegungen in Abschnitt „Technische Kriterien“.</p>	

- Werden die Lustre MetaDatenServer und die Lustre ObjektStorageServer als **separate Hardware** realisiert, sind **mindestens zwei** Server als **MetaDatenServer** und **mindestens zwei** weitere Server als **ObjektStorageServer** anzubieten.

<sup>23</sup> Siehe z.B. Hersteller DDN bzw. Celestica

- Werden die Lustre MetaDatenServer und ObjektStorageServer auf **derselben Serverhardware** realisiert, sind **mindestens zwei kombinierte MetaDaten/ObjektStorage – Server** anzubieten.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Lustre MetaDatenServer und ObjektStorageServer“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.5.3.3 Lustre MetaDatenTargets und ObjektStorageTargets

- Als MetaDatenTarget bzw. ObjektStorageTarget müssen zwischen den Lustre Servern **geteilte/shared Festplattensysteme/JBODs mit redundanten internen Controllern** eingesetzt werden.  
Die Möglichkeit, dass die nicht redundante Elektronik dieser Festplattensysteme ausfällt (Backplane Schaltkreise, SAS bzw. PCIe Expander u.s.w.) wird zugunsten der Wirtschaftlichkeit als Fehlerquelle akzeptiert. Das Kriterium der „kalten Redundanz“<sup>24</sup> gilt hier dennoch als erfüllt.
- Lustre MetaDatenTargets und ObjektStorageTargets müssen auf separaten Festplattensystemen laufen, d.h. ein **Mischbetrieb von MetaDatenTargets und ObjektStorageTargets** im selben Festplattensystem ist **nicht zulässig**.
- Um einen active / active Betrieb<sup>25</sup> der Lustre Server zu ermöglichen, müssen die oben genannten, geteilten Festplattensysteme logisch jeweils **in mindestens zwei** Lustre MetaDatenTargets bzw. ObjektStorageTargets unterteilt sein. Diese müssen **jeweils die gleiche Anzahl an SSDs bzw. HDDs** enthalten.
- Die Lustre MetaDatenTargets müssen **insgesamt mind. 160TB effektiv nutzbare Netto Kapazität**<sup>26</sup> **besitzen**.
- Die Lustre MetaDatenTargets müssen aus **insgesamt mind. 24 NVMe bzw. SAS SSDs** bestehen und mit **RAID Mirroring** abgesichert sein. Die Anbindung an den ObjektStorageServer muss über einen **herstellerspezifischen/proprietären PCIe Bus zum Sharen von NVMe SSDs zwischen zwei Servern/Nodes/Controllern**<sup>27</sup> oder **alternativ über SAS-4/SAS 24G** erfolgen. Das soll einer sehr hohen Performance und Ausfallsicherheit dienen.
- Die Übertragungsleistung der Lustre MetaDatenServer und MetaDatenTargets fließt in die Bewertung des Angebotes ein<sup>28</sup>.
- Die Lustre ObjektStorageTargets müssen **insgesamt mind. 1290TB effektiv nutzbare Netto Kapazität**<sup>29</sup> **besitzen**.
- Die Lustre ObjektStorageTargets müssen **HDD basiert** sein und über die Lustre Funktionalität **„Data on Metadata“** beschleunigt werden.
- Die Lustre ObjektStorageTargets müssen **insgesamt mind. 180 HDDs (CMR Verfahren, mind. 7200 upm)** umfassen und mit **RAID** abgesichert sein. Das soll einer hohen Performance und Ausfallsicherheit dienen.
- Die Übertragungsleistung der Lustre ObjektStorageServer und ObjektStorageTargets fließt in die Bewertung des Angebotes ein<sup>30</sup>.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Lustre MetaDatenTargets und ObjektStorageTargets“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

<sup>24</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>25</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>26</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>27</sup> Siehe z.B. Hersteller DDN bzw. Celestica

<sup>28</sup> Zu Details siehe Abschnitt „Zuschlagskriterien“

<sup>29</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>30</sup> Zu Details siehe Abschnitt „Zuschlagskriterien“

#### 3.5.3.4 NFS Server und NFS Backend Storage

- Es ist zulässig, die NFS Server und die Management Nodes auf **derselben Serverhardware** zu betreiben, wenn im Fehlerfall **mindestens ein weiterer kombinierter NFS/Management Server** zur Verfügung steht, der dessen Funktion übernimmt. Es muss eine **kalte Redundanz**<sup>31</sup> zwischen den NFS Servern gewährleistet sein.
- Als NFS Backend Storage muss ein **zwischen den NFS Servern geteiltes/shared Festplattensystem/JBOD mit redundanten internen Controllern** eingesetzt werden.  
Die Möglichkeit, dass die nicht redundante Elektronik der Festplattensysteme ausfällt (Backplane Schaltkreise, SAS bzw. PCIe Expander u.s.w.) wird zugunsten der Wirtschaftlichkeit als Fehlerquelle akzeptiert. Das Kriterium der „kalten Redundanz“<sup>32</sup> gilt hier dennoch als erfüllt.
- Der NFS Backend Storage muss auf einem separaten geteilten Festplattensystem/JBOD laufen, d.h. ein **Mischbetrieb von NFS Backend Storage und Managementdaten Storage** im selben Festplattensystem ist **nicht zulässig**.
- Der NFS Backend Storage muss **mind. 425TB effektiv nutzbare Netto Kapazität**<sup>33</sup> besitzen.
- Der NFS Backend Storage muss **mind. 60 HDDs (CMR Verfahren, mind. 7200 upm)** umfassen und mit **RAID** abgesichert sein. Die Anbindung an die NFS Server muss mind. über **SAS-4/SAS 24G** erfolgen. Das soll einer angemessenen Performance und hohen Ausfallsicherheit dienen.
- Der NFS Backend Storage muss für den Benchmark **io500.posix.iop-easy-write.nfs** einen Wert von **mindestens 4,8 GiB/s**<sup>34</sup> erreichen.
- Die Übertragungsleistung von NFS Server und NFS Backend Storage fließt in die Bewertung des Angebotes ein<sup>35</sup>.
- Werden die NFS Server und die Management Nodes als **separate Hardware** realisiert, sind **mindestens zwei Server als NFS Server** und **mindestens zwei Server als Management Node** anzubieten.
- Werden die NFS Server und die Management Nodes auf **derselben Serverhardware** realisiert, sind **mindestens zwei** kombinierte NFS/Management Server anzubieten.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „NFS Server und NFS Backend Storage“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

#### 3.5.3.5 HPC Managementdaten Storage

- Als Storage für die HPC Managementdaten muss ein **zwischen den Management Nodes geteiltes, SAS SSD basiertes Festplattensystem/JBOD mit redundanten Controllern** eingesetzt werden.  
Die Möglichkeit, dass die nicht redundante Elektronik der Festplattensysteme ausfällt (Backplane Schaltkreise, SAS Expander u.s.w.) wird zugunsten der Wirtschaftlichkeit als Fehlerquelle akzeptiert. Das Kriterium der „kalten Redundanz“<sup>36</sup> gilt hier dennoch als erfüllt.
- Der Storage für die Managementdaten muss auf einem separaten geteilten Festplattensystem laufen, d.h. ein **Mischbetrieb von Managementdaten Storage und NFS Backend Storage** im selben Festplattensystem ist nicht zulässig.
- Der Storage für die Managementdaten muss **mind. 6.900 GB effektiv nutzbare Netto Kapazität**<sup>37</sup> besitzen, aus **mind. 4 SAS SSDs** bestehen und über **RAID Mirroring** abgesichert sein.

<sup>31</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>32</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>33</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>34</sup> muss bei der Abnahme entsprechend den Vorgaben aus Abschnitt „Zuschlagskriterien - Benchmarks“ nachgewiesen werden

<sup>35</sup> Zu Details siehe Abschnitt „Zuschlagskriterien“

<sup>36</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>37</sup> Definition siehe Abschnitt „Technische Kriterien“

- Der Managementdaten Storage muss für den Benchmark **io500.posix.ior-easy-write.mgmt** einen Wert von **mindestens 5 GiB/s**<sup>38</sup> und für den Benchmark **io500.posix.mdtest-easy-write.mgmt** einen Wert von **mindestens 170 KIOPS** erreichen.
- Die Übertragungsleistung des Managementdaten Storage fließt in die Bewertung des Angebotes ein<sup>39</sup>.
- Sollten die Management Nodes und die NFS Server als **separate Hardware** realisiert werden, sind **mindestens 2 Server als Management Node** und **mindestens 2 Server als NFS Server** anzubieten.
- Sollten die Management Nodes und die NFS Server auf **derselben Serverhardware** realisiert werden, sind **mindestens 2 kombinierte Management/NFS – Server** anzubieten.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „HPC Managementdaten Storage“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.6 Netzwerk

#### 3.6.1 Netzwerk - Infiniband

- Es muss ein **Infiniband Netzwerk** zur Übertragung der HPC Daten zwischen den **Storage Servern** und allen **Compute-, Login- und Management Nodes** angeboten werden. Nur Infiniband verwendet hardwarebasiertes RDMA und kann MPI Operationen in den Adaptionen ausführen lassen. Das entlastet die CPUs der Server (Offload Architektur) und führt dadurch zu einem deutlich leistungsfähigerem Gesamtsystem als die Alternativen Omni-Path bzw. Ethernet.
- Alle Compute-, Login- und Management Nodes sowie alle Storage Server müssen mit **mind. NDR 200 Gbit/s** an das Infiniband Netzwerk angeschlossen sein.
- Es muss mindestens ein **unmanaged** Infiniband Switch mit **mind. 64 nutzbaren Ports zu je NDR 400 Gbit/s** bzw. **mind. 128 nutzbaren Ports zu je NDR 200 Gbit/s** angeboten werden. Das soll die nötige Performance sowie die zukünftige Erweiterbarkeit gewährleisten.
- Der angebotene Infiniband Switch muss mit den Ports/Konnektoren „zur Rückseite des Serverschrankes zeigend“ installiert werden. Deshalb ist ein Modell mit „**power-to-connector airflow**“ anzubieten.
- Der angebotene Infiniband Switch muss über **mind. zwei redundante Netzteile** verfügen.
- Alle **Infiniband Adapter** müssen **mind. NDR 200 Gbit/s** fähig sein sowie **GPUDirect RDMA** und **GPUDirect Storage** unterstützen. Letztere Funktionalitäten sollen den optimalen Betrieb der GPUs gewährleisten.
- Das Infiniband Netzwerk muss auf allen Netzwerkkomponenten (Switches, Netzwerkkarten im Betriebssystem) konfiguriert und funktionstüchtig übergeben werden.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Netzwerk - Infiniband“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

#### 3.6.2 Netzwerk - Ethernet

- Es muss ein Ethernet Netzwerk angeboten werden, welches in ein **Service- bzw. Administrationsnetzwerk** und ein **Managementnetzwerk** unterteilt ist. Über das Servicenetzwerk müssen u.a. die Daten des Ressourcen Managers übertragen und die Bootimages der Compute Nodes geladen werden. Im Managementnetzwerk müssen alle Daten zur Überwachung von Hard- und Software (BMC/IPMI) übertragen werden.

<sup>38</sup> muss bei der Abnahme entsprechend den Vorgaben aus Abschnitt „Zuschlagskriterien - Benchmarks“ nachgewiesen werden

<sup>39</sup> Zu Details siehe Abschnitt „Zuschlagskriterien“



- Service- und Managementnetzwerk müssen **segmentiert** sein, d.h. es darf **keine Layer-3 Verbindung/Routing** untereinander geben. Es ist eine physische oder eine logische Trennung über VLANs zugelassen.
- Das Ethernet Netzwerk muss **mind. 2 Switche** besitzen, die **untereinander mit mind. 2 Verbindungen zu je mind. 100Gbit verbunden** sind. Diese Switche werden im Folgenden als Core Switches bezeichnet, obwohl untypischerweise neben Switchen auch Nodes angeschlossen werden.
- Das Ethernet Netzwerk muss darüber hinaus **mind. einen weiteren Switch** (im Folgenden als Edge Switch bezeichnet) besitzen, der **redundant mit mind. zwei Verbindungen zu je mind. 10Gbit** an die Core Switches anzuschließen ist.
- Jeder angebotene Ethernet Switch muss **mind. zwei redundante Netzteile** besitzen.
- Jeder Compute Node muss mit **mind. ein Port zu mind. 10 GE** an das Servicenetzwerk über einen Core Switch angeschlossen sein. Die Compute Nodes der Bereiche CPU, GPU fp32 und GPU fp64 müssen **zu ungefähr gleichen Teilen auf die Core Switches verteilt** sein.
- Alle Login- und Management Nodes sowie Storage Server müssen **redundant mit mind. zwei Ports zu je mind. 10 GE** an das Servicenetzwerk über die Core Switches angeschlossen sein. Das beinhaltet, dass auf diesen Nodes eine **Failover - Software** betriebsbereit zu konfigurieren ist.
- Die Management Controller / BMCs aller Compute-, Login- und Management Nodes sowie die der Storage Server müssen mit **mind. einem Port zu mind. 1Gbit** an das Managementnetzwerk über einen Edge Switch angeschlossen sein.
- Die Managementschnittstellen aller HPC Infiniband- und Ethernet Switches, aller Festplattensysteme sowie von 14 Stromverteilerleisten müssen mit **mind. einem Port mit der maximalen Geschwindigkeit der jeweiligen Schnittstelle** an das Managementnetzwerk über einen Edge Switch angeschlossen sein.
- Die **Betriebssysteminstanzen der Management Nodes** (auf denen die Monitoringsoftware läuft) müssen redundant mit mind. zwei Ports zu mind. je 10GE an das Managementnetzwerk angeschlossen sein. Das beinhaltet, dass auf diesen Nodes eine **Failover - Software** betriebsbereit zu konfigurieren ist.
- Das Ethernet Netzwerk muss auf allen Netzwerkkomponenten (Switches, Netzwerkkarten usw.) **konfiguriert und funktionstüchtig** übergeben werden.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Netzwerk - Ethernet“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.7 HPC Software

#### 3.7.1 HPC Software - Allgemein

Jede im Folgenden erwähnte Software muss durch den Anbieter in der **höchsten aktuellen, stabilen Version** installiert und ggf. betriebsbereit konfiguriert werden. Zu den Details der anschließenden Konfiguration siehe in den folgenden Abschnitten.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „HPC Software - Allgemein“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

#### 3.7.2 Betriebssystem



- Auf allen Compute-, Management- und Login Nodes muss ein einheitliches **RHEL binärkompatibles und lizenzkostenfreies Linux** als Betriebssystem verwendet werden. Linux wird als Betriebssystem gefordert, weil es als Open Source Software kostenlos einsetzbar und unabhängig von Interessen privatwirtschaftlicher Hersteller ist. Weiterhin ist sämtliche in diesem Abschnitt beschriebene HPC Software unter Linux lauffähig, was für andere Betriebssysteme nur sehr eingeschränkt gilt.
- Auf den Management- und Login-Nodes muss das **Betriebssystem auf den lokalen Medien** installiert werden.
- Auf den Compute Nodes muss ein Betriebssystemimage über **UEFI HTTP bzw. über PXE gebootet** werden. Der Bootvorgang muss über **HPC Ethernet Servicenetzwerk** erfolgen.
- Auf den Management Nodes muss ein **Bootserver** installiert und betriebsbereit konfiguriert werden.
- Auf den Management- und Login-Nodes muss eine Software für die **Anbindung des Betriebssystems an die vorhandene Active Directory** installiert werden. Die weitere Konfiguration übernimmt die Auftraggeberin.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Betriebssystem“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.7.3 Dienste

- Als Containersystem muss Podman **installiert und betriebsbereit konfiguriert** werden.
- Jede Dienste - Software muss als **Quadlet Container oder über Paketmanager** auf dem jeweiligen Host installiert werden.
- Zwischen den Lustre MetaDatenServern, den Lustre ObjektStorageServern sowie den NFS Servern muss eine **Failover – Software** installiert und betriebsbereit konfiguriert werden. Die Lösung muss eine **Umschaltung der jeweiligen Storage Dienste zwischen den dazugehörigen Nodes** ermöglichen und dabei die Bedingungen einer **kalten Redundanz**<sup>40</sup> erfüllen.
- Zwischen den Management Nodes muss eine **Failover – Software** installiert und betriebsbereit konfiguriert werden. Die Lösung muss eine **Umschaltung der Management Dienste wie z.B. Provisionierung, Monitoring, Batchsystem usw. zwischen den Management Nodes** ermöglichen und dabei die Bedingungen einer **kalten Redundanz**<sup>41</sup> erfüllen.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Dienste“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

#### 3.7.3.1 Provisionierung

##### Für Management-, Login- und NFS Server

- Als Provisionierungssystem muss **Ansible** installiert und betriebsbereit konfiguriert werden.
- Die Auftraggeberin übernimmt die **Konfiguration von Ansible an das vorhandene Git System**.
- Für die Dienste Prometheus, Loki, Grafana und Alertmanager muss jeweils ein **Ansible Playbook** zur Verfügung gestellt werden, welches diese als **Quadlet Container** auf den Management Nodes **installiert**.
- Es muss ein **Ansible Playbook** zur Verfügung gestellt werden, welches **Warewulf** (warewulfd, dhcpd, tftp usw.) **über Paketmanager** auf den Management Nodes **installiert**.
- Es müssen **Ansible Playbooks** zur Verfügung gestellt werden, die die folgenden Dienste auf den Login- bzw. Management Nodes **installieren und betriebsbereit konfigurieren**:
  - **Ansible** (ein betriebsbereites Linux auf dem Zielsystem gilt als vorausgesetzt)

<sup>40</sup> Definition siehe Abschnitt „Technische Kriterien“

<sup>41</sup> Definition siehe Abschnitt „Technische Kriterien“

- **SSH** mit schlüsselbasiertem login vom Management zu den Login- bzw. Management Nodes
- **SLURM** (slurmd, slurmctld, MySql usw.)
- **MPI**
- **Infiniband Subnetmanager** auf Management Nodes
- **NFS Server**
- **Failover - Software**

#### Für Lustre Server

- Sollten die Lustre Server als „herkömmliche Server“, d.h. nicht als Appliance, ausgeführt sein, müssen **Ansible Playbooks** zur Verfügung gestellt werden, die die folgenden Dienste auf den Lustre Servern **installieren**:
  - **SSH** mit schlüsselbasiertem login vom Management zu den Lustre Servern
  - **Infiniband - Fabric**
  - **Lustre Software** für MDS, MDT, OSS, OST
  - **Failover Software**

#### Für Compute Nodes

- Für die Erstellung von Compute-Images muss **Warewulf** auf den Management Nodes **installiert** werden.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Provisionierung“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

#### 3.7.3.2 Monitoring

- Es muss eine HPC Managementlösung angeboten werden, die eine **Schnittstelle für Prometheus** besitzt.
- Die Software für **Prometheus, Loki, Grafana und Alertmanager** muss **als Quadlet Container auf den Management Nodes installiert** werden. Die weitere Konfiguration übernimmt die Auftraggeberin.
- Die Überwachung der Hardware muss über das **Redfish oder IPMI Protokoll** erfolgen.
- Störungsmeldungen müssen mindestens **über E-Mail** an die Mailadresse [hpc@itz.uni-halle.de](mailto:hpc@itz.uni-halle.de) übermittelt werden. Die Auftraggeberin stellt die Zugangsdaten zu ihrem Mailsystem.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Monitoring“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

#### 3.7.3.3 Batchsystem

- Als Batchsystem muss SLURM **installiert** und **betriebsbereit konfiguriert** werden. Folgende Slurm Build Einstellungen müssen aktiviert sein:
  - **nss\_slurm**
  - **configless**
  - **cgroupv2**
  - **Apptainer**
- Es muss das **Slurm Burst Buffer Plugin**<sup>42</sup> zur Verfügung gestellt werden, welches es den Anwendern ermöglicht, Daten vor der Ausführung von Jobs von Home nach Scratch und nach der Ausführung von Jobs von Scratch nach Home zu kopieren. Die weitere Konfiguration übernimmt die Auftraggeberin.

<sup>42</sup> Siehe [https://slurm.schedmd.com/burst\\_buffer.html](https://slurm.schedmd.com/burst_buffer.html)

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Batchsystem“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.7.4 Softwareerstellung

- Im Home Bereich muss ein **separates Dateisystem** für die Anwendungssoftware und die **Container Repositories** angelegt werden.
- Die im Folgenden erwähnte Software ist **nur in diesem Dateisystem und nicht im Linux Betriebssystem** zu installieren.
- Es muss das Modulsystem **Lmod installiert** und **betriebsbereit konfiguriert** werden.
- Es muss das Kompilierungstool **Spack installiert** und **betriebsbereit konfiguriert** werden.
- Folgende Software muss inklusive **Lua Lmod Modulen** bereitgestellt werden:
  - **GCC** mit C-, C++ und Fortran-Compilern (installiert via Spack)
  - Optimierte, an Infiniband angepasste **MPI-Variante**
  - **Intel oneAPI Base & HPC Toolkit**
  - **Nvidia HPC-SDK**
- Als Containersystem für HPC Anwendungen muss **Apptainer installiert** und **betriebsbereit konfiguriert** werden.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Softwareerstellung“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.8 Schulung und Dienstleistung

- Es sind **zwei Schulungen** zur Administration des HPC Systems über **jeweils 3 Tage**, also **insgesamt 6** Schulungstage anzubieten.
- An den Schulungen werden **3 Mitarbeiter der Auftraggeberin** teilnehmen.
- Die Schulungen müssen durch einen **qualifizierten deutschsprachigen HPC Ingenieur** in den **Räumen der Auftraggeberin** durchgeführt werden. Die Schulungsinfrastruktur stellt die Auftraggeberin.
- Ein Schulungstag umfasst **8 Arbeitsstunden ohne An- und Abreise**. Alle in Verbindung mit der Schulungsleistung anfallenden Kosten inklusive Schulungsunterlagen (in elektronischer Form) sind einem Preis anzubieten. Nachträglich berechnete Kosten sind ausgeschlossen.
- Die Schulungsunterlagen sind mind. 3 Werktage vor Beginn der Schulung den Teilnehmern zu übermitteln.
- Schulungsthemen sind **HPC Software, Message Passing Interface (MPI), die Dateisysteme Lustre und NFS, Konfiguration Infiniband und Ethernet, Server- und Storage – Hardware sowie frei wählbare Themen, die vorher mit dem Lehrenden abgestimmt werden**.
- Es sind **mind. 10 Tage Consulting / weiterführende Unterstützung** anzubieten. Das Consulting hat ein **qualifizierter deutschsprachiger HPC Ingenieur** durchzuführen. Die Ausführung muss über **Fernzugriff / per Remote** erfolgen.
- Ein Consultingtag umfasst 8 Arbeitsstunden. Alle in Verbindung mit der Consultingleistung anfallenden Kosten sind einem Preis anzubieten. Nachträglich berechnete Kosten sind ausgeschlossen.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Schulung und Dienstleistung“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

### 3.9 Vor - Ort Installation und Dokumentation

- Sämtliche in der Ausschreibung enthaltene Hardware muss in die von der Auftraggeberin bereitgestellten Serverschränke/Racks im Produktionsgebäude Kurt – Mothes - Straße 1, 06120 Halle **eingebaut, verkabelt und in einem betriebsbereiten Zustand** übergeben werden.
  - Sämtliche in der Ausschreibung enthaltene Software muss entsprechend den Abschnitten „Storage“ sowie „HPC Software“ **installiert, (wenn so gefordert) konfiguriert und in einem betriebsbereiten Zustand** übergeben werden.
  - Sämtliche in der Ausschreibung enthaltene Hard- und Software muss in **elektronischer** Weise in der Art dokumentiert werden, dass ein **Wiederaufbau des HPC Systems durch einen qualifizierten Dritten** anhand der Dokumentation möglich ist.
  - In der Dokumentation muss die **Struktur des HPC Systems** veranschaulicht sein (z.B. Anordnung der Komponenten im Rack, Belegungsplan Netzwerkverkabelung u.s.w.). Sämtliche **eingestellten Parameter** der Systeme sowie beispielhaft die **notigen Befehle zu deren Konfiguration** müssen hinterlegt sein.
  - Es muss eine **Vor-Ort Dokumentation** in Form von Beschriftung erfolgen. Konkret bedeutet das, dass **sämtliche Geräte und Kabel** beschriftet werden müssen. Die Bezeichnungen müssen denen in der Dokumentation entsprechen.
- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Vor - Ort Installation und Dokumentation“ beschriebenen Vorgaben im Angebot erfüllt wurden.**

## 4 Ausführungsbestimmungen

Die vereinbarten Lieferfristen sind verbindlich. Liefer- und Leistungsverzögerungen sind der MLU Halle Wittenberg unverzüglich anzuzeigen. Dies gilt dann, wenn es auf Grund von erheblichen und unvorhersehbaren Umständen politischen und/oder wirtschaftlichen Ausmaßes, zu Lieferengpässen bzw. Störungen in den Lieferketten kommt, die nicht im Verantwortungsbereich des Auftragnehmers liegen. Mit der zeitnahen Unterrichtung der MLU Halle-Wittenberg hierüber, sind die zu diesem Zeitpunkt voraussichtlichen Lieferzeiten mitzuteilen.

Alle in den Räumen der Auftraggeberin beschäftigten Mitarbeiter oder Erfüllungsgehilfen des Auftragnehmers müssen die sicherheitstechnischen Vorschriften der Auftraggeberin (Betriebsanweisung Maschinensaal - siehe Anlage) akzeptieren und deren Einhaltung schriftlich bestätigen. Darüber hinaus behält sich die Auftraggeberin eine Sicherheitsüberprüfung (einfach / SÜ1) dieser Personen gemäß Sicherheitsüberprüfungsgesetz vor.

Eventuell erforderliche Festanschlüsse von elektrischen Betriebsmitteln müssen von einer **geprüften Elektrofachkraft** erbracht werden.

Die Erstellung der Leistung untergliedert sich in die Ausführungsabschnitte:

1. **Lieferung der Hardware**
2. **Aufbau, Verkabelung und Konfiguration der Hardware**
3. **Installation und Konfiguration der Software bis zur Betriebsbereitschaft**
4. **Dokumentation**
5. **Schulung**

Jeder Ausführungsabschnitt muss durch ein Abnahmeprotokoll schriftlich durch das ITZ bestätigt werden. **Der Auftragnehmer kann für die tatsächlich erbrachten Leistungen für einen Ausführungsabschnitt eine Teilrechnung stellen.** Die Endabnahme erfolgt nach der Meldung der Betriebsbereitschaft durch den Auftragnehmer.

## 5 Zuschlagskriterien

### 5.1 Zuschlagskriterien - Allgemeines

Unter Berücksichtigung aller Umstände wird der Zuschlag auf das Angebot erteilt, das das beste Preis-Leistungsverhältnis erreicht. Die Bewertung erfolgt auf Grundlage des eingereichten Angebots. Daher liegt es im Interesse des Bieters, alle angeforderten Informationen so detailliert und korrekt wie möglich zur Verfügung zu stellen.

Bewertung der Angebote erfolgt nach der **erweiterten Richtwertmethode**. Als **Entscheidungskriterium** wird die **Zahl der Leistungspunkte** verwendet.

Als Preis wird der **Gesamtangebotspreis inkl. MwSt. in €, gerundet auf zwei Dezimalstellen hinter dem Komma** verwendet. Es gelten die kaufmännischen Rundungsregeln<sup>43</sup>.

Zur Ermittlung der Leistungspunkte werden folgende **Kriteriengruppen** verwendet, die im Abschnitt „Zuschlagskriterien - Benchmarks“ sowie in der angehangenen Bewertungsmatrix beschrieben werden:

1. Quantitative Kriterien
2. Qualitative Kriterien CPU
3. Qualitative Kriterien GPU
4. Qualitative Kriterien Storage
5. Lieferbedingungen
6. Garantie (**zusätzliche Lustre Software Garantie über 5 Jahre**)

Die Dauer zwischen Auftragserteilung und Endabnahme fließt in die Bewertung des Angebotes ein.<sup>44</sup>

### 5.2 Zuschlagskriterien - Benchmarks

#### 5.2.1 Rahmenbedingungen

Alle im **Angebot genannten Leistungswerte** müssen bei der Abnahme am betriebsbereiten HPC System vor Ort und entsprechend den Vorgaben im Abschnitt „Zuschlagskriterien - Benchmarks“ **nachgewiesen werden**. **Werden die im Angebot genannten Leistungswerte bei Abnahme nicht nachgewiesen, liegt kein mangelfreies Werk vor.**

Die **Ermittlung** aller im Angebot genannten Leistungswerte muss **transparent dokumentiert** sein, d.h. die Auftraggeberin muss nachvollziehen können, wie sie zustande gekommen sind.

Sollte für die Bestimmung der Leistungswerte eine vom angebotenen HPC System abweichende Testumgebung verwendet werden, muss deren **Aufbau und Konfiguration**, die darauf **ermittelten Werte** und der **Weg der Hochrechnung** auf das angebotene System dokumentiert werden. Weiterhin ist eine **Begründung für die Hochrechnungsmethode** anzufügen.

Es dürfen **nur ausdrücklich erlaubte Änderungen am Code der Benchmarks** vorgenommen werden. Es ist untersagt, Änderungen am Code oder den Eingabedaten vorzunehmen, die auf dem **Wissen um die Lösung des im Benchmark gestellten Problems** basieren.

Es ist untersagt, Code-Änderungen vorzunehmen, die dazu führen, dass die **Berechnung des Problems umgangen** wird.

Es sind **nur die vom Vendor unterstützten Compiler- und Linker-Flags** erlaubt.

<sup>43</sup> Definition siehe <https://de.wikipedia.org/wiki/Rundung>

<sup>44</sup> Zu Details siehe im Anhang „Bewertungsmatrix“

Es müssen **alle verwendeten Bibliotheken** angegeben werden. Das **Linken gegen optimierte Vendor-Bibliotheken** ist erlaubt.

Bei Verwendung von **Seeds** (für Zufallszahlengeneratoren o.ä.) müssen diese mit dokumentiert werden. Die **BIOS-Einstellungen** müssen vor Beginn der Benchmarks gesetzt werden und bis zum Ende sämtlicher Messungen unverändert bleiben.

**Hyperthreading** sowie **Turbo Boost (Intel)/Turbo Core (AMD)** müssen disabled sein.

Jede Änderung am verwendeten **Algorithmus eines Benchmarks** muss vollständig offengelegt werden und bedarf der Genehmigung durch die Auftraggeberin.

Alle Benchmarks müssen auf dem Scratch/Lustre Dateisystem laufen, außer es ist explizit anders angegeben.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Rahmenbedingungen“ beschriebenen Vorgaben bei der Ermittlung der Benchmarkwerte erfüllt wurden.**

## 5.2.2 CPU Benchmarks

Zur Bewertung der Rechenleistung der Compute Nodes werden die Benchmarks HPCC und HPCG verwendet.

### 5.2.2.1 HPCC - Erlaubte und nicht erlaubte Änderungen

Erlaubt sind:

- Compiler- und Linker-Flags, die vom Vendor unterstützt werden und dokumentiert sind
- Linken gegen optimierte Vendor-Versionen folgender Bibliotheken: BLAS, FFT, MPI  
Hinweis : Alle so verwendeten Bibliotheken müssen beim Einreichen der Ergebnisse mit angegeben werden  
Hinweis : Aufrufe von Bibliotheks-Subroutinen müssen die gleiche Syntax und Semantik haben wie im veröffentlichten Code des Benchmarks. Veränderungen am Benchmark-Code mit dem Zweck, die Aufrufe an die verwendeten Vendor-Bibliotheken anzupassen, sind nicht zulässig.

Nicht erlaubt sind:

- Verwendung von GPU-Beschleunigern
- Änderungen der Genauigkeit: Berechnungen müssen mit doppelter Fließkommagenauigkeit (64-bit, double precision) durchgeführt werden.

### 5.2.2.2 HPCC - Kompilierung

Es werden benötigt:

- C-Compiler
- MPI-Bibliothek
- BLAS-Bibliothek
- FFT-Bibliothek

Laden Sie die HPCC Quellcode von folgender Webseite herunter und entpacken Sie das Archiv:

[https://hpc.itz.uni-halle.de/ausschreibung\\_2025/hpcc-1.5.0.tar.gz](https://hpc.itz.uni-halle.de/ausschreibung_2025/hpcc-1.5.0.tar.gz)

Folgen Sie den Anweisungen in der Datei README.txt, die im Hauptverzeichnis des Archivs liegt:

- Erstellen Sie eine Datei setup/Makefile.<arch> und passen Sie diese entsprechend an. Beachten Sie dabei die erlaubten und nicht erlaubten Anpassungen.
- Führen Sie make <arch> aus

Das Ergebnis ist eine Binärdatei hpcc im HPCC-Hauptverzeichnis.

### 5.2.2.3 HPCC - Durchführung

Der HPCC Benchmark muss auf **1536 Cores der Compute Node CPUs** durchgeführt werden.

Setzen Sie die Konfigurationsoptionen für HPCC in der Datei **hpccinf.txt** im HPCC-Hauptverzeichnis.

Hinweis: Um die beste Leistung des angebotenen Systems zu ermitteln, sollte die größte Problemgröße verwendet werden, die in den Hauptspeicher passt. Die von HPL verwendete Speicher-  
menge entspricht im Wesentlichen der Größe der Koeffizientenmatrix.

Der Benchmark HPCC kann z.B. mit „mpirun -np 4 ./hpcc“ gestartet werden.

Die exakten Startoptionen sind hierbei abhängig von der verwendeten MPI-Implementierung und dem Ressourcen-Manager.

### 5.2.2.4 HPCC - Einzureichende Werte

Um die Ermittlung der Werte nachvollziehen zu können, ist neben den im Folgenden genannten Benchmark - Ergebnissen auch die zugrundeliegende Datei **hpccoutf.txt** mit dem Angebot einzureichen.

#### hpcc.randomAccess

Der einzureichende Wert steht in Datei **hpccoutf.txt** in der **Sektion „StarRandomAccess“** und dort in der **Zeile, die mit "Average GUP/s" beginnt**. Die Maßeinheit ist GUP/s.

#### hpcc.stream.copy, hpcc.stream.scale, hpcc.stream.add, hpcc.stream.triad

Die vier einzureichenden Werte stehen in Datei **hpccoutf.txt** in der **Sektion „StarSTREAM“** und dort in den **Zeilen, die mit „Average Copy“, „Average Scale“, „Average Add“ sowie „AverageTriad“ beginnen**. Die Maßeinheit ist GB/s.

#### hpcc.ptrans

Der einzureichende Wert „PTRANS“ steht in Datei **hpccoutf.txt** in der **Sektion „PTRANS“**. Dort ist der **größte Wert aus der Spalte „GB/s“** zu ermitteln. Die Maßeinheit ist GB/s.

#### hpcc.pingPong.latency, hpcc.pingPong.bandwith

Die zwei einzureichenden Werte stehen in Datei **hpccoutf.txt** in der **Sektion „LatencyBandwidth“**. Dort sind im **Unterabschnitt „Ping Pong“** aus den **Zeilen, die mit „Latency“ bzw. „Bandwith“ beginnen**, die Werte aus der **Spalte „avg“** zu ermitteln. Die Maßeinheit für Latency ist msec, die für Bandwith ist MByte/s.

#### hpcc.naturalRing.latency, hpcc.naturalRing.bandwith

Die zwei einzureichenden Werte stehen in Datei **hpccoutf.txt** in der **Sektion „LatencyBandwidth“**. Dort sind im **Unterabschnitt „Ring“** aus den **Zeilen, die mit „On naturally“ beginnen**, die Werte **rechts von „latency=“ und von „bandwith=“** zu ermitteln. Die Maßeinheit für Latency ist msec, die für Bandwith ist MByte/s.

#### hpcc.randomRing.latency, hpcc.randomRing.bandwith

Die zwei einzureichenden Werte stehen in Datei **hpccoutf.txt** in der **Sektion „LatencyBandwidth“**. Dort sind im **Unterabschnitt „Ring“** aus den **Zeilen, die mit „On randomly“ beginnen**, die Werte **rechts von „latency=“ und von „bandwith=“** zu ermitteln. Die Maßeinheit für Latency ist msec, die für Bandwith ist MByte/s.

### 5.2.2.5 HPCG - Erlaubte und nicht erlaubte Änderungen



Erlaubt sind:

- Anpassen der Eingabedatei hpcg.dat (siehe Hinweise in "Durchführung" zur Auswahl einer angemessenen Problemgröße)
- Compiler- und Linker-Flags, die vom Vendor unterstützt werden und dokumentiert sind
- Linken gegen optimierte Vendor-Versionen folgender Bibliotheken: BLAS, MPI  
Hinweis: Alle so verwendeten Bibliotheken müssen beim Einreichen der Ergebnisse mit angegeben werden  
Hinweis: Aufrufe von Bibliotheks-Subroutinen müssen die gleiche Syntax und Semantik haben wie im veröffentlichten Code des Benchmarks. Veränderungen am Benchmark-Code mit dem Zweck, die Aufrufe an die verwendeten Vendor-Bibliotheken anzupassen, sind nicht zulässig.

Nicht erlaubt sind:

- Verwendung von GPU-Beschleunigern
- Eine Vektorengröße, die kleiner als 25% des verwendeten Hauptspeichers ist
- Eine Vektorengröße, die so klein ist, dass sie in den Third Level Cache einer CPU passt (wird berechnet durch  $n_x * n_y * n_z * \text{sizeof}(\text{double})$ )
- Eine Laufzeit < 1800 Sekunden
- Änderungen der Genauigkeit: Berechnungen müssen mit doppelter Fließkommagenauigkeit (64-bit, double precision) durchgeführt werden.

#### 5.2.2.6 HPCG - Kompilierung

Es werden benötigt:

- C-Compiler
- MPI-Bibliothek
- BLAS-Bibliothek

Laden Sie die HPCG Quellcode von folgender Webseite herunter und entpacken Sie das Archiv:

[https://hpc.itz.uni-halle.de/ausschreibung\\_2025/hpcg-114602d.tar.gz](https://hpc.itz.uni-halle.de/ausschreibung_2025/hpcg-114602d.tar.gz)

Folgen Sie den Anweisungen in der Datei INSTALL, die im Hauptverzeichnis des Archivs liegt:

- Erstellen Sie eine Datei setup/Makefile.<arch> und passen Sie diese entsprechend an. Beachten Sie dabei die erlaubten und nicht erlaubten Anpassungen.
- Erstellen Sie ein build-Verzeichnis und wechseln Sie dort hinein
- Führen Sie ./configure <arch> aus

Das Ergebnis sind die zwei Binärdateien bin/xhpcg sowie bin/hpcg.dat im build-Verzeichnis.

#### 5.2.2.7 HPCG - Durchführung

Der HPCG Benchmark muss auf **1536 Cores der Compute Node CPUs** durchgeführt werden.

Setzen Sie die Konfigurationsoptionen für HPCG in der Datei **hpcg.dat**, die nach der Kompilierung im Verzeichnis bin neben der Benchmark-Binary xhpcg liegt.

Hinweis: In Zeile 3 können Sie die Größe eines rank-lokalen Vektors anpassen.

Die Größe des Vektors wird berechnet durch  $n_x * n_y * n_z * \text{sizeof}(\text{double})$ , wobei die Parameter  $n_x$ ,  $n_y$  und  $n_z$  den drei Zahlen in der Zeile 3 entsprechen und  $\text{sizeof}(\text{double})$  die Größe des C-Datentyps double auf der von Ihnen angebotenen Plattform ist (im Normalfall 8 Bytes).

Achten Sie darauf, dass für eine gültige Einreichung die Vektorengröße sowohl **größer als der Third Level Cache einer CPU** als auch **größer als 25% des verwendeten Hauptspeichers** sein muss.

Hinweis: In Zeile 4 legen Sie die Laufzeit des Benchmarks fest.

Für gültige Einreichungen müssen Sie hier einen **Wert >= 1800** benutzen.



Der Benchmark HPCG kann z.B. mit „mpirun -np 4 ./bin/xhpcg“ gestartet werden  
Die exakten Startoptionen sind hierbei abhängig von der verwendeten MPI-Implementierung und dem Ressourcen-Manager.

#### 5.2.2.8 HPCG - Einzureichende Werte

Um die Ermittlung der Werte nachvollziehen zu können, ist neben den im Folgenden genannten Benchmark - Ergebnissen auch die zugrundeliegende Datei **HPCG-Benchmark\_3.1\_<timestamp>.txt** mit dem Angebot einzureichen.

##### hpcg

Der einzureichende Wert steht in der Datei **HPCG-Benchmark\_3.1\_<timestamp>.txt** in der **Zeile, die mit Final Summary beginnt**. Dort ist der Wert rechts von „GFLOP/s rating of=“ zu ermitteln. Die Maßeinheit ist GFLOP/s.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „CPU Benchmarks“ beschriebenen Vorgaben bei der Ermittlung der Benchmarkwerte erfüllt wurden.**

#### 5.2.3 GPU Benchmarks

Zur Bewertung der Übertragungsleistung zwischen zwei GPUs wird der Benchmark **osu\_bibw** verwendet.

##### 5.2.3.1 osu\_bibw - Erlaubte und nicht erlaubte Änderungen

Erlaubt sind:

- Änderung der Optionen für den Launcher (siehe Abschnitt „osu\_bibw - Durchführung“).
- Änderung an der Umgebungsvariablen CUDA\_VISIBLE\_DEVICES
- Änderung des Mappings von MPI - Prozessen auf Cores
- Änderung des Skripts get\_local\_rank

##### 5.2.3.2 osu\_bibw - Kompilierung

Es werden benötigt:

- funktionierende MPI-Implementierung
- CUDA-SDK

Laden Sie den Code der OSU Micro Benchmarks von folgender Webseite herunter und entpacken Sie das Archiv:

[https://hpc.itz.uni-halle.de/ausschreibung\\_2025/osu-micro-benchmarks-7.5.tar.gz](https://hpc.itz.uni-halle.de/ausschreibung_2025/osu-micro-benchmarks-7.5.tar.gz)

Der Benchmark benutzt das Autotools - Buildsystem und wird mit „configure && make && make install“ kompiliert. Es ist zu beachten, dass die Umgebungsvariablen CC und CXX zwingend gesetzt werden müssen.

Das Ergebnis ist die Binärdatei „./c/mpi/pt2pt/standard/osu\_bibw“.

##### 5.2.3.3 osu\_bibw - Durchführung

Der Benchmark ermittelt die Bandbreiten zwischen zwei GPU-Puffern. Beim Benchmark werden GPUs **innerhalb eines Nodes** und auf **zwei unterschiedlichen Nodes** getestet. Weiterhin sind die GPUs der **GPU fp32** sowie der **GPU fp64** Compute Nodes zu testen. Es sind also vier Läufe nötig.

Der Benchmark `osu_bibw` kann z.B. mit „**mpirun osu\_bibw -z -d cuda D D**“ gestartet werden. Die exakten Startoptionen sind hierbei abhängig von der verwendeten MPI-Implementierung und dem Ressourcen-Manager.

Der MPI-Launcher ist so zu konfigurieren, dass:

- zwei MPI-Prozesse **auf einem Node** gestartet werden, die jeweils Zugriff auf eine GPU haben (für den Benchmark **osu.bibw.intra-node**)
- je ein MPI-Prozess auf zwei **unterschiedlichen Nodes** gestartet wird und diese MPI-Prozesse Zugriff auf je eine GPU haben (für den Benchmark **osu.bibw.inter-node**).

Bitte beachten Sie, dass die GPU-Affinität für Prozesse noch vor der MPI-Initialisierung festgelegt wird. Der `osu_bibw` Benchmark wertet dafür die Umgebungsvariable `LOCAL_RANK` aus. Bitte exportieren Sie die node-lokale MPI-Rank ID in die Umgebungsvariable `LOCAL_RANK`.

Ein Beispiel für die Ermittlung von **osu.bibw.intra-node**:

```
„mpirun -n 2 ./get_local_rank ./c/mpi/pt2pt/standard/osu_bibw -z -d cuda D D“
```

Ein Beispiel für die Ermittlung von **osu.bibw.inter-node**:

```
„mpirun -n 2 -N 1 --hostfile hostfile ./get_local_rank ./c/mpi/pt2pt/standard/osu_bibw -z -d cuda D D“
```

#### 5.2.3.4 osu\_bibw - Einzureichende Werte

Um die Ermittlung der Werte nachvollziehen zu können, ist neben den im Folgenden genannten Benchmark - Ergebnissen auch die zugrundeliegende **vollständige Konsolenausgabe** mit dem Angebot einzureichen.

##### osu.bibw.intra-node.fp32

Dieser Benchmark ist **auf zwei GPUs eines GPU fp32 Compute Nodes** laufen zu lassen.

Der einzureichende Wert ist der höchste in **Spalte Bandwith** erzielte Wert **der Konsolenausgabe**. Die Maßeinheit ist MB/s.

##### osu.bibw.inter-node.fp32

Dieser Benchmark ist **auf jeweils einer GPU zweier GPU fp32 Compute Nodes** laufen zu lassen.

Der einzureichende Wert ist der höchste in **Spalte Bandwith** erzielte Wert **der Konsolenausgabe**. Die Maßeinheit ist MB/s.

##### osu.bibw.intra-node.fp64

Dieser Benchmark ist **auf zwei GPUs eines GPU fp64 Compute Nodes** laufen zu lassen.

Der einzureichende Wert ist der höchste in **Spalte Bandwith** erzielte Wert **der Konsolenausgabe**. Die Maßeinheit ist MB/s.

##### osu.bibw.inter-node.fp64

Dieser Benchmark ist **auf jeweils einer GPUs zweier GPU fp64 Compute Nodes** laufen zu lassen.

Der einzureichende Wert ist der höchste in **Spalte Bandwith** erzielte Wert **der Konsolenausgabe**. Die Maßeinheit ist MB/s.

☐ **Der Anbieter bestätigt, dass alle im Abschnitt „GPU Benchmarks“ beschriebenen Vorgaben bei der Ermittlung der Benchmarkwerte erfüllt wurden.**

#### 5.2.4 Storage Benchmarks

Zur Leistungsbewertung der vier Storage Bereiche **Scratch/Lustre, Home/NFS Storage, HPC Management Storage** und **Lokale SSDs Compute Nodes** wird der IO500 Benchmark verwendet.

#### 5.2.4.1 IO500 - Zusätzliche Rahmenbedingungen

Die Regeln sind für unsere Zwecke leicht abgewandelt von <https://io500.org/rules/submission>.

- Es gelten die grundlegenden Regeln aus dem Abschnitt "Rahmenbedingungen".
- Das Stonewall-Flag muss auf **300** gesetzt werden.
- Die Dateinamen für die MDTest- und IOR-Ausgabedateien dürfen **nicht vorab angelegt** sein.
- Für jedes zu testende Speichersystem (Lustre, NFS, lokale SSDs) müssen alle Benchmark - Phasen **ohne Unterbrechung** durchgeführt werden.
- Keine der zu testenden Benchmark-Phasen darf in der result\_summary.txt-Ergebnisdatei als **[INVALID]** gekennzeichnet sein.
- Es dürfen nur Datenblöcke im Benchmark - Ergebnis erfasst werden, die **vollständig auf das Testmedium geschrieben** wurden. Das bedeutet umgekehrt, dass alle Datenblöcke, die sich zum Messungsende noch im IO Cache befinden, nicht gewertet werden dürfen.
- Daten und Metadaten müssen **vollständig auf das Testmedium übertragen** werden, d.h. Datenreduzierung über Deduplizierung, Komprimierung usw. ist untersagt.
- Wenn für die Suchphase ein anderes „find“ Tool als das enthaltene pfind verwendet wird, muss es **dasselbe Eingabe- und Ausgabeverhalten** wie pfind aufweisen. Der Quellcode des verwendeten Tools muss der Auftraggeberin zur Verfügung gestellt werden.
- Die 10 Node Benchmark - Messungen müssen auf **10 Compute Nodes gleichzeitig ausgeführt werden**. Auf jedem Node muss **mindestens je ein Benchmark - Prozess** ausgeführt werden.
- Für jede der vier Hauptphasen ior-easy, ior-hard, mdtest-easy und mdtest-hard darf **vorab ein Verzeichnis erstellt und angepasst werden**. Weitere Unterverzeichnisse dürfen nicht vorab erstellt werden.

#### 5.2.4.2 IO500 - Erlaubte und nicht erlaubte Änderungen

Erlaubt sind:

- Änderungen der Anzahl der MPI-Prozesse in der Datei **io500.sh**:
- Änderungen an den Variablen io500\_ini, io500\_mpirun und io500\_mpiargs
- Änderungen an der Funktion setup()

Nicht erlaubt sind:

- Änderungen, die gegen die „Rahmenbedingungen“ oder die „zusätzlichen Rahmenbedingungen“ für diesen Benchmark verstoßen.

folgende Einstellungen in der Datei **config.ini**

- drop-caches = FALSE
- scc = TRUE
- Ein kleinerer Wert als stonewall-time = 300
- Alle weiteren Änderungen in der config.ini

#### 5.2.4.3 IO500 - Kompilierung

- Installieren Sie eine MPI-Implementierung
- Laden Sie die IO500 Quellcode von folgender Webseite herunter und entpacken Sie das Archiv: [https://hpc.itz.uni-halle.de/ausschreibung\\_2025/io500-sc24.tar.gz](https://hpc.itz.uni-halle.de/ausschreibung_2025/io500-sc24.tar.gz)
- Wechseln Sie in den entsprechenden Ordner und führen Sie **./prepare.sh** aus

#### 5.2.4.4 IO500 - Durchführung

- Erzeugen Sie eine Datei config.ini mit „**io500 --list > config.ini**“
- Passen Sie die Datei **config.ini** entsprechend des Laufes an

- Passen Sie in der Datei **setup.sh** die Funktion `setup()` und die Variablen **io500\_ini**, **io500\_mpi-run** und **io500\_mpiargs** an.
- Starten Sie den Benchmark mit „./io500.sh“

Je nach Festlegung im Abschnitt „IO500 - Einzureichende Werte“ müssen folgende Phasen **auf einem und gegebenenfalls zusätzlich auf 10 Nodes** getestet werden:

Mit der POSIX API:

- `ior-easy-{write,read}`
- `ior-hard-{write,read}`
- `mdtest-easy-{write,stat,delete}`
- `mdtest-hard-{write,stat,read,delete}`
- `find`

Mit der MPI-IO API:

- `ior-easy-{write,read}`
- `ior-hard-{write,read}`

Für jede Kombination aus API (POSIX oder MPI-IO) und Anzahl an Nodes (1 oder 10) muss in ein **separater Benchmark - Lauf** durchgeführt werden.

#### 5.2.4.5 IO500 - Einzureichende Werte

Um die Ermittlung der Werte nachvollziehen zu können, ist neben den im Folgenden genannten Benchmark - Ergebnissen auch das jeweils zugrundeliegende **vollständige Ergebnisverzeichnis** mit dem Angebot einzureichen. Die Ergebnisverzeichnisse sind diejenigen, die die Dateien **results.txt** und **result\_summary.txt** enthalten.

Für den Storagebereich **Scratch/Lustre** sind einzureichen :

Für folgende mit „io500.posix“ beginnende Werte muss **ein Benchmark – Lauf auf einem und einer auf zehn Compute Nodes** (also insgesamt zwei) durchgeführt werden.

##### io500.posix.ior-hard-read.1node und io500.posix.ior-hard-read.10nodes

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „ior-hard-read“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist GiB/s.

##### io500.posix.ior-hard-write.1node und io500.posix.ior-hard-write.10nodes

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „ior-hard-write“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist GiB/s.

##### io500.posix.mdtest-hard-read.1node und io500.posix.mdtest-hard-read.10nodes

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-hard-read“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

##### io500.posix.mdtest-hard-write.1node und io500.posix.mdtest-hard-write.10nodes

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-hard-write“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

##### io500.posix.mdtest-hard-stat.1node und io500.posix.mdtest-hard-stat.10nodes

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-hard-stat“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

**io500.posix.mdtest-hard-delete.1node und io500.posix.mdtest-hard-delete.10nodes**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-hard-delete“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

**io500.posix.find.1node**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „find“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

Für folgende mit „io500.mpio“ beginnende Werte muss **ein Benchmark – Lauf auf zehn Compute Nodes** durchgeführt werden.

**io500.mpio.ior-hard-read.10nodes**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „ior-hard-read“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist GiB/s.

**io500.mpio.ior-hard-write.10nodes**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „ior-hard-write“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist GiB/s.

Für den Storagebereich **Home/NFS** sind einzureichen:

Für die folgenden Werte von „Home/NFS“ muss **ein Benchmark - Lauf auf einem Login Server auf dem gemounteten NFS Share von Home** durchgeführt werden. Der Grund dafür ist, dass nur die Login Server NFS Schreibberechtigung haben dürfen.

**io500.posix.ior-easy-read.nfs**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „ior-easy-read“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist GiB/s.

**io500.posix.ior-easy-write.nfs**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „ior-easy-write“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist GiB/s.

Für den Storagebereich **HPC Managementdaten** sind einzureichen:

Für die folgenden Werte von „HPC Managementdaten“ muss **ein Benchmark - Lauf auf dem aktiven Management - Node auf dem Management Storage** durchgeführt werden.

**io500.posix.ior-easy-read.mgmt**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „ior-easy-read“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist GiB/s.

**io500.posix.ior-easy-write.mgmt**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „ior-easy-write“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist GiB/s.

**io500.posix.mdtest-easy-write.mgmt**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-easy-write“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

**io500.posix.mdtest-easy-stat.mgmt**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-easy-stat“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

**io500.posix.mdtest-easy-delete.mgmt**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-easy-delete“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

Für den Storagebereich **Lokale SSDs Compute Nodes** sind einzureichen:

Für die folgenden Werte von „Lokale SSDs Compute Nodes“ muss **ein Benchmark - Lauf auf einem Compute Node auf der lokalen SSD** durchgeführt werden.

**io500.posix.mdtest-hard-read.ssd**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-hard-read“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

**io500.posix.mdtest-hard-write.ssd**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-hard-write“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

**io500.posix.mdtest-hard-stat.ssd**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-hard-stat“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

**io500.posix.mdtest-hard-delete.ssd**

Der einzureichende Wert steht in der Datei **result\_summary.txt** in der **Zeile mit „mdtest-hard-delete“** und befindet sich rechts von diesem Schlüsselwort. Die Maßeinheit ist KIOPS.

- ☐ **Der Anbieter bestätigt, dass alle im Abschnitt „Storage Benchmarks“ beschriebenen Vorgaben bei der Ermittlung der Benchmarkwerte erfüllt wurden.**

### **5.3 Zuschlagskriterien - Bewertungsmatrix**

Die Bewertungsmatrix ist den Ausschreibungsunterlagen als separate Anlage angefügt. Sie enthält ein Blatt zur Ermittlung des Gesamtpreises, eines zur Ermittlung der Werte für die Bewertungskriterien und eines zur Ermittlung der Leistungspunkte.