

Leistungsbeschreibung

„Erweiterung eines GPU-Server-Clusters“

Die RPTU Kaiserslautern-Landau betreibt am Standort Kaiserslautern im Fachbereich Informatik einen GPU-Server-Cluster, der um zwei weitere leistungsfähige Deep-Learning-Server erweitert werden soll.

Dieser Einsatzzweck erfordert daher für Hochleistungsrechnen im Dauerbetrieb („24/7“) – insbesondere für das längere Training von KI-Modellen - ausgelegte Hardware. Dies betrifft besonders die Kühlung der aktiven Komponenten (Mainboard, CPU, GPU, RAM, Storage). Die Server werden in bauseitig bereitgestellte 19“-Racks in einem speziellen GPU-Server-Raum mit Luftkühlung (kaltwassergestützte Umluftkühlung über rackseitig angereicherte Luft/Wasser-Wärmetauscher am bauseitigen Kaltwassernetz) eingebaut. Das angebotene System muss luftgekühlt sein (Front-to-Back-Luftführung), eine Direktwasserkühlung ist nicht vorgesehen.

Alle Angaben des Angebots sind technisch detailliert zu gestalten, um die Kompatibilität und Integrationsfähigkeit in das vorhandene GPU-Server-Cluster sicherstellen zu können.

Nebenangebote und Änderungsvorschläge sind nicht zugelassen.

Der Einbau durch den Auftragnehmer umfasst den Einbau der Hardware in bauseitig bereitgestellte 19“-Racks.

Nach erfolgtem Einbau durch den Auftragnehmer werden die Systeme durch den Auftraggeber einer Abnahme-Prozedur inklusive eines 3-tägigen Dauertests unterzogen. Die Systeme gelten als abgenommen, wenn dieser Dauertest ohne Ausfälle und Leistungseinbußen erfolgreich abgeschlossen wurde.

Den Zuschlag erhält das günstigste Angebot.

Liefer- und Einbauadresse:
RPTU Kaiserslautern-Landau
Gebäude 48, Raum 259
Gottlieb-Daimler-Straße
67663 Kaiserslautern

Lieferung und Einbau zweier Deep-Learning-Server

Geliefert und installiert werden sollen 2 Serversysteme vom Typ

„NVIDIA DGX B300-288GB“ oder gleichwertig

mit jeweils 8 Blackwell Ultra GPUs (B300 288GB) und 3 Jahren Wartung.

Das Budget für diese Beschaffungsmaßnahme liegt bei ca. 800.000 EUR inkl. MwSt.

Folgende Systemeigenschaften müssen durch die 2 **Deep-Learning-Server** jeweils erfüllt werden:

- GPUs:
 - 8x NVIDIA B300 Tensor Core GPUs
 - verbunden über 2x NVIDIA® NVSwitch™ (keine reine PCIe-GPU-Topologie)
- GPU-Hauptspeicher:
 - 2304 GB total
- Stromversorgung:
 - maximal 15 kW
 - 230V AC / 50 Hz
 - Anschluss an C19/C20 Rack-PDUs
- CPU:
 - Mindestens 2x 32 Cores, x64 kompatibel
- System Hauptspeicher:
 - 2 TB
- Netzwerk:
 - mindestens 1 Gb/s onboard NIC mit RJ45
 - Host baseboard management controller (BMC) mit RJ45
 - 8 OSFP-Ports über 8 ConnectX-8-Karten oder gleichwertige Hochgeschwindigkeits-Netzwerkadapter mit 8 × 800 Gb/s InfiniBand/Ethernet
 - 2 Dual-Port BlueField-3-DPUs oder gleichwertige DPU-/SmartNIC-Adapter mit 400 Gb/s InfiniBand/Ethernet
- Storage:
 - OS: mindestens 2x 1 TB NVMe
 - Daten: mindestens 8x 3 TB NVMe
- Systemsoftware:
 - Betriebssystem: NVIDIA DGX OS 7 oder gleichwertiges Linux-Betriebssystem mit Long Term Support, bevorzugt Ubuntu-LTS-basiert
 - Vorinstallierter und zertifizierter Standard-KI-Software-Stack – insbesondere:
GPU-Treiber und GPU-Computing-Laufzeitumgebung für angebotene GPU-Generation,
CUDA als GPU-Programmier- und Laufzeitumgebung,
Container-Laufzeitumgebung einschließlich GPU-Container-Unterstützung,
Werkzeuge für Systemmanagement, GPU-/System-Monitoring, Health Checks und Fehlerdiagnose,
vom Hersteller validierte Treiber und Managementwerkzeuge für die angebotenen Hochgeschwindigkeits-Netzwerkadapter,
Möglichkeit zur Einspielung von sicherheits- und funktionsrelevanten Herstellerupdates während des Supportzeitraums.

Der Einbau der beiden Deep-Learning-Server erfolgt durch den Auftragnehmer in vorhandene 19“-Racks. Die Herstellung der Betriebsbereitschaft umfasst das Booten des vorinstallierten Betriebssystems und die Demonstration der Funktionsfähigkeit aller Systemkomponenten.