Leistungsbeschreibung „Hochskaliertes 6G KI Backbone für große Sprachmodelle“ 2026-101

Anforderungen

Zielsetzung/Kurzbeschreibung
Der **Lehrstuhl für Kommunikationsnetze der Technischen Universität Dortmund** führt Forschungsarbeiten im Bereich der **6G-Mobilfunktechnologien** durch. Für die Simulation, Analyse und Entwicklung von Verfahren der künstlichen Intelligenz (KI) im Kontext von 6G-Netzen wird ein **Hochskaliertes 6G KI-Backbone für große Sprachmodelle** benötigt, das als zentrale Rechenressource für datenintensive Trainings- und Inferenzprozesse sowie Simulationen dient. Das System soll eine hochintegrierte **vorkonfigurierte, wissenschaftlich einsetzbare KI-Hochleistungsplattform** bereitstellen, die speziell für das Deep-Learning bzw. Foundation-Modell-Training, modellgestützte Netzwerksimulationen und Echtzeitdatenverarbeitung konzipiert ist.
Technische Anforderungen / Formfaktor
Die Deep-Learning-Serverplattform zur parallelen Datenverarbeitung und Modellberechnung besteht aus einer vollintegrierten Basisplattform mit folgender Spezifikation: ○ Zwei x86_64-kompatible Prozessoren mit jeweils min. 50 physischen Rechenkernen (100 Threads) und min. 2GHz Basistakt ○ Unterstützung von PCIe 5.0 und DDR5-Arbeitsspeicher ○ Ausstattung mit min. 2 TB DDR5 Arbeitsspeicher ○ Systemlaufwerke mit min. 2 × 1,9 TB NVMe SSD für das Betriebssystem ○ Zusätzlicher lokaler Speicher gesamt ≥ 30 TB NVMe (U.2/U.3) ○ Luftkühlung (Strömungsrichtung vorne nach hinten) ○ Bauweise in 19-Zoll Serverrack-Format, Bauhöhe max. 16 Höheneinheiten ○ KI-Rechenleistung im Training min. 70 PetaFLOPS (FP8, Sparse) bzw. 140 PetaFLOPS (FP4, Sparse) bei Inferenz ○ Unterstützung der Datentypen FP64, TF32, BF16, FP16, FP8, FP4, INT8 ○ Zwei zusätzliche Dual-Port-Netzwerkadapter mit Bluefield-3 DPU mit je min. 100 GbE ○ Bereitstellung von mindestens 3 Netzwerk-Schnittstellen: ▪ 10 Gb/s RJ45 Onboard-Netzwerkschnittstelle ▪ 100 Gb/s Dual-Port-Netzwerkadapter (s.o.) ▪ Baseboard-Management-Controller mit dediziertem RJ45-Port Die enthaltenen grafischen Recheneinheiten (GPUs) weisen die folgenden Spezifikationen auf: ○ CUDA-Unterstützung mit CUDA Compute-Capability >= 9.0 zur Unterstützung von spezialisierter 6G- sowie Robotik-KI-Software von NVIDIA ○ Gesamter verfügbarer GPU-Speicher: min. 1400 GB mit HBM3e Technologie ○ Interconnect-Architektur mit einer bidirektionalen Bandbreite zwischen zwei beliebigen GPUs ≥ 1800 GB/s und einer Gesamtbandbreite ≥ 14 TB/s (non-blocking) ○ Speicherbandbreite zu lokalem GPU-Speicher pro GPU ≥ 7 TB/s Anforderungen an das Systemmanagement und bereitgestellte Software:

- Lieferung als vollständig betriebsbereite Einheit, inklusive installiertem Basisbetriebssystem und Managementsoftware
- Native Unterstützung von Container-basierten Workflows (z. B. Docker, Podman, Kubernetes)
- Fernwartungsfähige Managementschnittstelle (BMC) mit Monitoring- und Kontrollfunktionen
- Bereitstellung betriebsfertiger Container-Images für verbreitete Deep-Learning-Frameworks (inkl. PyTorch, TensorFlow, Theano)

- Sonstiges:
  - Spannungsversorgung über 230V AC / 50 Hz, Anschlusskabel im Lieferumfang zum Anschluss an C19/C20 Rack-PDUs
  - Redundanz mit Ausfallsicherheit min. eines Netzteils
  - Maximale Gesamtleistung des Systems < 15kW
  - Netzteile mit mindestens 80 PLUS Titanium Zertifikation
  - Min. 3 Jahre Hardware Wartung und Vor-Ort-Service durch den Hersteller
  - **Lieferung inkl. Montage vor Ort** *(Anlieferung siehe Anhang)*
  - **Vollständige Einrichtung des Systems sowie Einweisung in die Nutzung**

Sonstiges

- Angebotsobergrenze 290.527,- € (Netto, zzgl. Mehrwertsteuer)