

Leistungsbeschreibung Los 1

Erweiterung der NVIDIA DGX GPU-Infrastruktur — 4× DGX B300

Projekt: Naton x RACOON |

1. Ausgangssituation und Zielsetzung

Das Universitätsklinikum Essen (UK Essen) betreibt im Rahmen eines Forschungsprojekts eine bestehende NVIDIA-DGX-GPU-Infrastruktur für rechenintensive KI-Anwendungen in der medizinischen Bildanalyse und klinischen Forschung.

Zur Deckung des steigenden Rechenbedarfs für Training, Fine-Tuning und Inferenz komplexer KI-Modelle (insbesondere im Bereich Radiomics, Segmentierung und multimodale Analyse) soll die vorhandene Infrastruktur um **vier (4) NVIDIA DGX B300 Systeme** erweitert werden.

Die Lieferung muss den in dieser Leistungsbeschreibung unter Punkt 3 genannten Fabrikaten, Modellbezeichnungen, Stückzahlen und Einzelkonfigurationen **exakt entsprechen**. Sämtliche Geräte sind fabrikneu und in der beschriebenen Konfiguration zu liefern.

1.1 Projektziele

- Erhöhung der verfügbaren GPU-Rechenkapazität um 32× NVIDIA Blackwell Ultra B300 GPUs
- Nahtlose Integration in die bestehende DGX-/Cluster-Umgebung
- Sicherstellung des Betriebs über 5 Jahre durch Hersteller-Support
- Schnelle Inbetriebnahme (Time-to-Production) durch schlüsselfertige Lieferung und Installation
-

1.2 Einsatzbereich

Anwendungsfall	Beschreibung
Deep Learning Training	Training großer Modelle für medizinische Bildanalyse
Fine-Tuning	Anpassung vortrainierter Modelle an klinische Datensätze
Inferenz	Echtzeit- und Batch-Inferenz für Forschungs-Pipelines

Anwendungsfall	Beschreibung
Radiomics	Feature-Extraktion und Analyse aus Bilddaten
Pipelines	Integration in bestehende Forschungs-Workflows

2. Leistungsumfang

Der Auftragnehmer erbringt folgende Leistungen:

2.1 Lieferung

- Lieferung von **4 fabrikneuen NVIDIA DGX B300 Systemen** als Hersteller-Appliance
- Inklusive aller im Leistungsverzeichnis genannten Hardwarekomponenten
- Inklusive **5-jähriger NVIDIA-Hersteller-Hardwarewartung** und Vor-Ort-Service
- Vorinstalliertes DGX OS and DGX Software Stack
- Transport **frei Haus** bis zum Installationsort (Rechenzentrum UK Essen)

2.2 Installation

- Mechanische Rack-Montage aller 4 Systeme (je 10 HE) in vorbereiteten 19"-Racks
- Anschluss an die vorhandene Stromversorgung (n+n Redundanz)
- Verkabelung der Stromanschlüsse (C19/C20)
- Einhaltung aller NVIDIA-Installationsanforderungen (Abstand, Kühlung, Zuluft)

2.3 Inbetriebnahme

- Funktionsprüfung aller Systeme (POST, GPU-Erkennung, Storage, BMC/Redfish)
- Start und Validierung des DGX OS und der inkludierten DGX Software Stack
- Basis-Konfiguration für Integration in die Cluster-Umgebung
- Verifikation: 8x B300 GPU pro System vollständig erkannt und funktionsfähig
- Dokumentation aller Seriennummern und Konfigurationsparameter
- Einweisung des IT-Betriebspersonals (mindestens 2 Stunden vor Ort)
- Schlüsselfertige Übergabe mit unterzeichnetem Übergabeprotokoll

2.4 Wartung und Support

- 5 Jahre NVIDIA Hardware-Support ab Abnahme
 - Remote-Diagnose und Vor-Ort-Austausch defekter Komponenten
 - Erweiterte RMA (Return Merchandise Authorization), Next-Business-Day
 - Firmware- und BIOS-Updates im Rahmen des Supportvertrags
-
-

3. Technische Anforderungen

Jedes der 4 Systeme muss folgende Spezifikation erfüllen:

3.1 Compute

Komponente	Anforderung
GPU	8× NVIDIA Blackwell Ultra B300 Tensor-Core-GPUs (SXM)
GPU-Speicher	2.304 GB gesamt (288 GB pro GPU, HBM3e)
GPU-Interconnect	NVIDIA NVSwitch, 5. Generation NVLink
Training Performance	72 PFLOPS (FP8)
Inferenz Performance	144 PFLOPS (FP4)
CPU	2× Intel Xeon Platinum 6776P, 128 Kerne gesamt
RAM	2 TB DDR5 (erweiterbar auf 4 TB)

3.2 Storage

Komponente	Anforderung
Boot	2× 1,92 TB M.2 NVMe
Cache/Data	30 TB (8× 3,84 TB E1.S NVMe, Self-Encrypting)

3.3 Netzwerk (Host-seitig, ohne Fabric)

Komponente	Anforderung
Cluster-NICs	8× OSFP mit je 1× ConnectX-8 VPI, bis 800 Gb/s IB/Ethernet
DPU	2× dual-port BlueField-3 DPU mit ConnectX-8, bis 400 Gb/s IB/Ethernet
Management	1 GbE RJ-45 (BMC), IPMI/Redfish

3.4 System

Komponente	Anforderung
OS	DGX OS (Ubuntu-basiert), vorinstalliert
Software Stack	NVIDIA DGX Software Stack, vorinstalliert
Formfaktor	10 HE (10U), 19"-Rack, Tiefe ≤ 920 mm
Gewicht	170 kg
Strom	12× 3,2 kW PSU, n+n; max. 15,1 kW; 12× C19/C20 Kabel
Kühlung	Luftgekühlt, Zuluft 5–30 °C
Support	5 Jahre NVIDIA Hardware-Support inkl. Vor-Ort-Service

4. Standort und bauliche Voraussetzungen (AG)

Der Auftraggeber stellt bereit:

Voraussetzung	Spezifikation
Rack-Platz	4× 10 HE frei, Einbautiefe ≥ 920 mm
Strom	n+n, ausreichend für 4× max. 15,1 kW (ca. 60 kW gesamt)
Kühlung	Ausreichend für 4× max. 15,1 kW Wärmeabgabe
Zutritt	Zugang Rechenzentrum, Anlieferungsroute
Netzwerk-Patchfelder	Anschlusspunkte am Rack (passiv, ohne aktive Komponenten)
Ansprechpartner	IT-Betrieb vor Ort während Installation

5. Abnahmekriterien

Die Abnahme erfolgt nach erfolgreicher Erbringung aller Leistungen gemäß folgender Checkliste:

- ☐ 4 Systeme physisch installiert und verkabelt
- ☐ Alle Seriennummern dokumentiert
- ☐ POST erfolgreich auf allen Systemen
- ☐ 8× B300 GPU pro System erkannt (nvidia-smi)
- ☐ DGX OS und Software Stack funktionsfähig
- ☐ BMC/Redfish erreichbar
- ☐ Storage vollständig erkannt
- ☐ Support-Vertrag für 5 Jahre dokumentiert
- ☐ Einweisung durchgeführt und protokolliert
- ☐ Übergabeprotokoll unterzeichnet