BALTIC BRAIN / COMPUTE

Fraunhofer IGD

Technische Spezifikation

Hardware, Plattform und Nutzungsrahmen für FuE-Projekte

Baltic Brain ist die KI-Forschungsinfrastruktur des Fraunhofer IGD für Training, Modellentwicklung, Evaluation und Inferenz. Für Ihr Vorhaben vereinbaren wir passende GPU- und Speicherkontingente sowie die benötigte technische Unterstützung.

Cluster und Standort

Merkmal	Ausstattung
Systeme und GPUs	3 DGX-B200-Nodes; je 8 NVIDIA B200, insgesamt 24 GPUs
SSD-Speicher	400 TB
Standort	Green IT Cube von GSI/FAIR in Darmstadt, Deutschland
Nutzung und Betreuung	Fraunhofer IGD Rostock und Darmstadt; in Betrieb seit August 2026

Technische Plattform je DGX-B200-Node

Komponente	Plattformspezifikation
GPU-Speicher	192 GB HBM3e je GPU; 1.536 GB je Node, insgesamt 4.608 GB verteilt auf die 24 GPUs
GPU-Kopplung	NVLink der 5. Generation; 2 NVSwitches je Node
CPU	2 Intel Xeon Platinum 8570, je 56 Kerne; 2,1 GHz Basis, bis 4,0 GHz Turbo
Arbeitsspeicher	2 TB je Node
Cluster-Netzwerk	4 OSFP-Anschlüsse für 8 ConnectX-7-VPI-Adapter mit bis zu 400 Gb/s je Adapter; InfiniBand/Ethernet
Storage-/Datennetz	2 BlueField-3-DPUs mit 400-Gb/s-Anbindung
Lokaler NVMe-Speicher	Betriebssystem: 2 × 1,92 TB; Datencache: 8 × 3,84 TB = 30,72 TB je Node
Bauform / Leistung	10 Höheneinheiten; Leistungsaufnahme bis ca. 14,3 kW je System

Kapazität passend zu Ihrem Workload

Sie können einzelne GPUs, mehrere GPUs innerhalb eines Nodes oder projektbezogene Kontingente nutzen. CPU, Arbeitsspeicher und Datenablagen stimmen wir auf Ihren Workload ab. Der Zugriff auf mehrere GPUs und deren Speicher erfolgt über entsprechend ausgelegte Trainings- und Inferenzverfahren.

BALTIC BRAIN / COMPUTE

Software, Schnittstellen und Modelle

Software und Dienste

Baustein	Funktion
NVIDIA Run:ai	Projektbezogene Ressourcenverwaltung; Training, Inferenz und Workspaces; volle GPUs oder geteilte Ressourcen
Workload-Übergabe	Containerbasierte Ausführung; Weboberfläche und CLI, unterstützte Kubernetes-/Framework-Manifeste per YAML
LLM-Inferenz	vLLM-Workloads auf dem Cluster; Modellgewichte über persistenten Cache (PVC)
API-Gateway	LiteLLM mit OpenAI-kompatibler API, persönlichen API-Keys sowie Nutzungs- und Budget-Tracking
Chat-Oberfläche	LibreChat für Dialoge mit Sprachmodellen; RAG, MCP-Werkzeuge und Diagramme nach projektbezogener Vereinbarung
Monitoring	Grafana für Modelle und Cluster; GPU-Belegung, Speichernutzung und Token-Metriken
Projektspeicher	S3-Speicher auf der Cluster-NetApp; Speicherquote und weitere Datenablagen nach Vereinbarung

API-Nutzung und eigene Workloads

Der API-Zugang zu bereitgestellten Modellen erfolgt über einen **persönlichen API-Key**; ein eigenes Run:ai-Projekt ist dafür nicht erforderlich. Für eigene Trainings- und Inferenzworkloads erhalten Sie eine **Projekt- und Ressourcenzuweisung**. Die Einrichtung erfolgt auf Grundlage des vereinbarten Sicherheits- und Zugriffsmodells.

Modellübersicht

Modell, Version, Nutzungsrechte und Kapazität werden für Ihr Vorhaben vereinbart. Die Kontextgrößen geben die maximal konfigurierte Länge an; Durchsatz und mögliche Parallelität hängen von Modell, Anfrageumfang und zugewiesenen Ressourcen ab.

Rolle	Modell	Kontext / Einordnung
`standard`	GLM-5.3-Flash	1.048.576 Token; Text und Bild
`premium`	GLM-5.3	262.144 Token; Text, Reasoning und Coding
`fast`	Qwen3.8-27B FP8	1.010.000 Token; Testbetrieb
Spezialist	Gemma 4 26B-A4B	131.072 Token; multimodale Anwendungen
Spezialist	Ornith 1.5 35B	131.072 Token; Coding und Tool-Aufrufe

Integration und technische Begleitung

Die OpenAI-kompatible Schnittstelle erleichtert die Anbindung eigener Anwendungen und vorhandener Werkzeuge. Wir unterstützen Sie bei Modellauswahl, Datenvorbereitung, Workload-Auslegung und Evaluation. Gemeinsam legen wir fest, welche Plattformfunktionen Ihr Projekt benötigt und wie sie in Ihre Arbeitsabläufe eingebunden werden.

BALTIC BRAIN / COMPUTE

Nutzung, Abrechnung und Projektgrenzen

Zugang und Datenkontrolle

Externer Zugriff wird erst nach Freigabe eines verbindlichen Sicherheits- und Zugriffsmodells durch Datenschutz und IT-Sicherheit bereitgestellt. Es regelt benannte Nutzerkonten, Authentisierung, Rollen und Berechtigungen, Mandantentrennung, zulässige Datenablagen und Protokollierung.

Speicherort, Zugriffsrechte, Datenverarbeitung, Löschung und Rückgabe sowie technische und organisatorische Maßnahmen werden vor Projektbeginn festgelegt. Rechte an Daten, Modellen und Ergebnissen sind Bestandteil der Projektvereinbarung. Für optionale Werkzeuge und Integrationen sind die jeweiligen Datenflüsse gesondert zu berücksichtigen.

Abrechnung im FuE-Auftrag

Die Abrechnung erfolgt je nach vereinbarter Leistung **pro GPU-Stunde oder tokenbasiert**. Alle Compute-Leistungen werden ausschließlich im Rahmen eines vertraglich definierten Forschungs- und Entwicklungsauftrags **nach unseren AGB** erbracht. API- und GPU-Nutzung müssen einen wesentlichen Beitrag zum FuE-Vorhaben leisten.

Vereinbarung	Inhalt
Arbeitsplan	Forschungsfrage, Arbeitspakete, Ergebnisse, Verantwortlichkeiten und Laufzeit
GPU-Abrechnung	GPU-Anzahl bzw. Anteil, abrechenbarer Zeitraum, Taktung und Kontingent; 8 volle GPUs für 1 Stunde ergeben 8 GPU-Stunden
Token-Abrechnung	Modell und Version, Zählweise, Input-/Output-Token, ggf. Cache-Token, Preise und Limits
Weitere Leistungen	Speicher, Betreuung, Datenaufbereitung und Entwicklung; Einbeziehung in den Preis oder gesonderte Positionen
Betriebsbedingungen	Verfügbarkeit, Reaktionszeiten, Wartung, Backup-/Wiederherstellungsumfang und Support

Abgrenzung

Kein eigenständiges Cloud-, Hosting- oder GPU-Vermietungsangebot. Kein Bare-Metal-Zugriff und keine uneingeschränkte Systemadministration durch Kunden. Ein hochverfügbares kommerzielles Service-Level wird nicht zugesagt. Ein späterer Produktivbetrieb beim Kunden wird gesondert geplant.

AI-Act-Beratung ist eine getrennte Leistung. Sie ist weder Bestandteil eines GPU-/Token-Tarifs noch Voraussetzung für die Beauftragung von Compute. Die technische Workload-Beratung gehört zum vereinbarten Compute-Leistungsumfang.

Technische Projektplanung: Sprechen wir über Ihren Workload.

Matthias Vahl | Fraunhofer IGD, Rostock
matthias.vahl@igd-r.fraunhofer.de | +49 381 4024-157
www.igd.fraunhofer.de