Der Informatiklehrstuhl VIII für Künstliche Intelligenz plant die Erweiterung seiner GPU-Server-Ressourcen durch die Anschaffung eines Servers mit vier NVIDIA H200 Grafikprozessoren. Damit soll den stetig wachsenden Anforderungen an Rechenkapazität im Bereich des maschinellen Lernens und der künstlichen Intelligenz begegnet werden.

Die NVIDIA H200-GPUs verfügen über besonders großen Speicher (141 GB HBM3e pro Karte) und hohe Speicherbandbreite, wodurch umfangreiche Modelle effizient trainiert und ausgeführt werden können. Ein zentraler Vorteil liegt in der Multi-Instance-GPU-Technologie (MIG), die eine flexible Aufteilung jeder GPU in mehrere unabhängige Einheiten ermöglicht. Dadurch kann der Lehrstuhl seine Ressourcen dynamisch an wechselnde Anforderungen anpassen – von großskaligen Trainingsläufen bis hin zu vielen kleineren Experimenten und studentischen Projekten. In der Praxis erlaubt dies eine Skalierung zwischen vier sehr großen Modellen und bis zu 112 kleineren Instanzen.

Diese Flexibilität ist insbesondere in Lehre und betreuter Forschung entscheidend. Studierende und Mitarbeitende können Arbeitsumgebungen nutzen, die exakt an ihre Aufgaben angepasst sind – etwa kleinere, ressourcenschonende Konfigurationen für Lehrveranstaltungen, Projektarbeiten und Abschlussarbeiten. Damit wird eine verlässliche Bereitstellung der benötigten Rechenumgebung gewährleistet, unabhängig von der Auslastung zentraler Systeme.

Zwar steht mit LIDO4 eine universitätsweite GPU-Infrastruktur zur Verfügung, jedoch ist sie für kurzfristige, interaktive oder experimentelle Vorhaben nur bedingt geeignet, da der Betrieb auf stabile Konfigurationen und gemeinsame Nutzung vieler Lehrstühle ausgerichtet ist. Eine Änderung der GPU-Aufteilung (MIG) würde bei jeder Neukonfiguration ein bis zwei LIDO4-Einheiten mit mehreren GPUs vorübergehend stoppen. Für Lehrveranstaltungen oder parallele Projekte stünden nicht ausreichend GPUs zeitnah bereit. Der eigene Server ermöglicht hier eine flexible, unmittelbar steuerbare Nutzung, bei der Konfigurationsänderungen ausschließlich den Lehrstuhl betreffen und keine Auswirkungen auf andere Nutzer haben.

Darüber hinaus trägt die Anschaffung zur effizienten Verwendung universitärer Ressourcen bei. Kurze, kleinteilige Trainings- und Testläufe – wie sie in Lehr- und Entwicklungsprojekten häufig vorkommen – würden auf LIDO4 eine ganze GPU belegen und umfangreichere Forschungsrechnungen anderer Gruppen blockieren. Durch die eigenständige Nutzung am Lehrstuhl können solche kleineren Aufgaben ausgelagert werden, wodurch das zentrale System gezielt für große, rechenintensive Aufträge entlastet wird. Die Anschaffung ergänzt LIDO4 somit funktional, anstatt in Konkurrenz dazu zu treten.

Ein weiterer Vorteil liegt in der Möglichkeit, mit den geteilten MIG-Einheiten experimentelle Szenarien zu simulieren, die sich an zukünftigen Anwendungsumgebungen orientieren – etwa an Rechenressourcen von autonomen Fahrzeugen, lokalen Servern oder mobilen Endgeräten. Eine aufgeteilte GPU-Instanz entspricht dabei typischen Hardware-Beschränkungen solcher Systeme und erlaubt realitätsnahe Forschung, die auf zentralen Systemen nicht effizient abbildbar wäre.

Mit der Anschaffung wird die Rechenkapazität des Lehrstuhls gezielt, wirtschaftlich und zukunftssicher erweitert. Forschende, Lehrende und Studierende erhalten damit die notwendige Flexibilität, um moderne KI-Modelle effizient zu entwickeln, zu trainieren und praxisnah einzusetzen.

Auf Dieser Basis wurden folgende konkrete Leistungsanforderungen identifiziert, welche der anzuschaffende GPU-Server erfüllen sollte:

System:	Gigabyte oder Supermicro
CPU:	2x AMD EPYC 9655
RAM:	2304GB = 24 x 96GB (mindestens 6400 Mt/s)
GPU:	4x NVIDIA H200 NVL 141 GB (EDU), Erweiterbar bis 8x NVIDIA H200
Front-Bay:	mind. 4x HotSwap U.2/3 und SATA/SAS fähige Front Bays
Drives:	4x U.2/3 NVME SSD, ~7,6TB, 1-2 DWPD, in Front
	1x M.2 NVMe SSD, ~960 GB, 1-2 DWPD, on MB
Netzwerk:	2x Ports, mind. 10GbE
BMC:	lizenzfreier BMC/IPMI, RJ45
Chassis:	4-6 HE
Netzteil:	redundante Platinum Netzteile