

Einweihung des Supercomputers Hunter am HLRS

Das HLRS hat den Supercomputer Hunter am 16. Januar feierlich eingeweiht. Das von Hewlett Packard Enterprise entwickelte System kombiniert hohe Rechenleistung mit energieeffizienter Technologie. Hunter wird u. a. für Simulationen, KI und Datenanalysen in Wissenschaft und Industrie eingesetzt. Mit einer theoretischen Spitzenleistung von 48,1 PetaFlop/s und einer innovativen Architektur, die CPUs und AMD Instinct™ Accelerated Processing Units integriert, bietet er deutliche Fortschritte in Leistung und Nachhaltigkeit. Hunter setzt auf direkte Flüssigkeitskühlung und ein dynamisches Power-Capping-System zur Optimierung des Energieverbrauchs. Im Vergleich zum Vorgängersystem Hawk reduziert Hunter den Energiebedarf um 80%. Die Kombination von CPUs und GPUs in einem einzigen Paket wird auch die Entwicklung und Ausführung neuartiger Converged-Computing-Workflows erleichtern, die Simulation, Datenanalyse und KI kombinieren. Das 15-Mio.-Euro-Projekt wurde zu gleichen Teilen vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg sowie vom BMBF (SiVeGCS-Projekt) finanziert. Die Landesministerin Petra Olschowski unterstrich die Bedeutung von HPC und KI für Wissenschaft und Wirtschaft. Weitere Informationen: hlrs.de/de/news/detail/hlrs-feiert-die-einweihung-des-supercomputers-hunter. (Kontakt: [Sophia Honisch](mailto:Sophia.Honisch@HLRS@GCS), HLRS@GCS)

Blue Lion: ein neuer Supercomputer fürs LRZ

Das LRZ hat Hewlett Packard Enterprise mit dem Aufbau seines nächsten Höchstleistungsrechners beauftragt. Blue Lion wird ab 2027 als GCS-System für nationale Spitzenforschung klassische HPC-Workflows sowie KI-Methoden ausführen und kombinieren. Dafür nutzt Blue Lion die nächste Generation von HPE-Cray-Technologie sowie NVIDIA-Prozessoren und -Beschleunigern. Für den schnellen Datentransfer sorgt das Interconnect HPE Slingshot mit einer Übertragungsrate von 400 Gigabit Daten pro Sekunde. Gekühlt wird das System direkt und zu 100% mit Wasser. Derart ausgerüstet bewältigt Blue Lion ungefähr das 30-fache der Rechenleistung des Vorgängersystems SuperMUC-NG. Das Zusammenspiel von CPU-Kernen und GPU-Beschleunigern erfordert neues Programmieren. Um die Anwender:innen für die effiziente Nutzung des neuen Supercomputer vorzubereiten, bietet das LRZ bereits ab diesem Jahr Workshops an. Die Gesamtkosten von 250 Mio. Euro für Blue Lion teilen sich das BMBF sowie das Bayerische Staatsministerium für Wissenschaft und Kunst. Weitere Informationen: tiny.badw.de/gtjJss. (Kontakt: presse@lrz.de, LRZ@GCS)

HammerHAI: AI Factory für Wissenschaft und Industrie

Das EuroHPC-Projekt HammerHAI wird in Deutschland eine AI Factory aufbauen. Koordiniert vom HLRS sind außerdem die GA-Partnerzentren LRZ, GWDG und KIT, sowie die SICOS BW GmbH für Simulation, Computing und Speicher beteiligt. HammerHAI wird eine sichere, in Deutschland ansässige Plattform und einen Datenspeicher bereitstellen. Ziel ist es, Forschung, Industrie und öffentliche Einrichtungen mit einer sicheren und KI-optimierten HPC-Infrastruktur zu unterstützen. Der Zugriff auf die entstehenden KI-Modelle sowie die ihnen zugrundeliegenden Daten und die Nutzungsschnittstellen werden entsprechend streng kontrolliert. Dies wird es europäischen Forschenden und Unternehmen erleichtern, die Compliance-Anforderungen für die Datensicherheit und ethische Nutzung von KI einzuhalten. Mit einem Gesamtbudget von rund 85 Mio. Euro wird das Projekt von der EuroHPC Joint Undertaking sowie mehreren deutschen Ministerien kofinanziert. Es ist Teil der europaweiten Strategie, leistungsfähige KI-Infrastrukturen bereitzustellen und Innovationen voranzutreiben. Weitere Informationen: hlrs.de/de/news/detail/hammerhai-wird-eine-ai-factory-fuer-wissenschaft-und-industrie-aufbauen-1. (Kontakt: [Sophia Honisch](mailto:Sophia.Honisch@HLRS@GCS), HLRS@GCS)

HOSHMAND – KI-gesteuerter Compute Scheduler

Das HOSHMAND-Framework (High-performance Open-sourced AI-based Scheduling Handler for Managing Node Distribution) stellt einen KI-gesteuerten Scheduler vor, der die Verteilung von Rechenjobs auf HPC-Systemen beschleunigt und die Ressourcenauslastung verbessert. Diese an der Universität Göttingen entwickelte Lösung nutzt ein rekurrentes neuronales Netzwerk (RNN), um die optimale Job-Knoten-Zuordnung vorherzusagen, und historische Daten, um Strategien dynamisch zu verfeinern. HOSHMAND passt sich schnell an schwankende Arbeitslasten und unterschiedliche Rechenanforderungen an. Seine Fähigkeit, mit heterogenen Ressourcen umzugehen, gewährleistet eine zuverlässige Leistung in Cloud-Umgebungen mit unterschiedlichen Hardware-Konfigurationen. Erste Tests mit dem Datensatz MIT SuperCloud haben gezeigt, dass die Planungszeiten reduziert und der Durchsatz verbessert werden konnten. Dieses Projekt ist eine Zusammenarbeit mit der DECICE-Initiative (Device-Edge-Cloud Intelligent Collaboration Framework, Beteiligung GWDG), die vom BMBF gefördert wird. Weitere Informationen: gwdg.de/about-us/gwdg-news/2024/GN_11-2024_www.pdf. (Kontakt: [Michael B. Khani](mailto:Michael.B.Khani@GWDG), GWDG)

Erste deutschsprachige große Sprachmodelle

Mit dem erfolgreichen Training von gleich zwei rein deutschen großen Sprachmodellen setzt der Lehrstuhl für Data Science am Center for Artificial Intelligence and Data Science (CAIDAS) der Universität Würzburg einen Meilenstein für deutschsprachige Large Language Models (LLMs). Beide Modelle – „LLäMmlein 120M“ und das leistungsstärkere „LLäMmlein 1B“ mit einer Mrd. Parametern sowie verschiedenen Chat-Varianten – stehen Forschenden seit November 2024 zur freien Verfügung. Das Projekt, bei dem nicht nur die deutsche Sprachverarbeitung, sondern auch die Untersuchung und Verbesserung der Lerndynamik der Modelle im Fokus steht, ist Auftakt für die Entwicklung noch größerer Modelle. Die umfangreichen Berechnungen für das 1B-Modell wurden am GPGPU-Cluster „Alex“ des NHR@FAU in Erlangen durchgeführt. Hierfür wurden 50.000 GPU-Stunden auf A100-GPUs mit 80 GB Speicher benötigt. Das kleinere Modell wurde auf dem neuen eigenen JuliaV2-Cluster der Universität Würzburg mit etwa 10.000 L40-GPU-Stunden berechnet. Weitere Informationen: informatik.uni-wuerzburg.de/datascience/projects/nlp/llammlein/. (Kontakt: [Andreas Hotho](mailto:Andreas.Hotho@jmu.de), JMU)

Teleportation in VR: Projekt „Put Me There“

Virtual Reality (VR) ermöglicht neue Interaktionsformen in immersiven 3D-Umgebungen, die in der echten Welt nicht möglich sind. Mit diesem Fokus will das DFG-Projekt „Put Me There“ das Potenzial von VR besser für die Wissenschaft erschließen. Neben Spielen wird VR bereits vielseitig zu Schulungszwecken oder zur medizinischen Unterstützung eingesetzt. Eine Herausforderung ist die Navigation in größeren virtuellen Räumen, wo Teleportation als bevorzugte Methode gilt, um sensorische Diskrepanzen und Cyberkrankheit zu vermeiden. Das Projekt der Partner RWTH Aachen und Universität Trier untersucht Teleportationstechniken und deren Nutzungsfreundlichkeit. Es analysiert Zielgenauigkeit und entwickelt neue Ansätze, um zusätzliche Freiheitsgrade zu integrieren. Dazu gehören das Teleportieren in den Luftraum über Objekten und die simulierte Skalierung von Nutzer:innen im virtuellen Raum. Ziel ist es, die Nutzererfahrung zu verbessern und innovative Perspektiven innerhalb von VR-Systemen zu ermöglichen. Weitere Informationen: blog.rwth-aachen.de/itc/2024/11/13/put-me-there/. (Kontakt: [Tim Weißker](mailto:Tim.Weißker@itc.rwth-aachen.de), IT Center)

QUADRIGA – Datenkompetenz für Berlin-Brandenburg

Das Datenkompetenzzentrum QUADRIGA vereint die vier Disziplinen Digital Humanities, Verwaltungswissenschaft, Informatik und Informationswissenschaft entlang der Datentypen Text, Tabelle und Bewegtes Bild am Wissenschaftsstandort Berlin-Brandenburg. QUADRIGA schafft Begegnungsorte zur Erfassung, Aufbereitung, Nutzung und Analyse von Daten in einer zunehmend di-

gitalisierten Welt, um Wissenschaftler:innen aller Karrierestufen Ressourcen für den souveränen Umgang mit Daten zur Verfügung zu stellen. Zur Entwicklung digitaler Methoden werden mittels Fallstudien in den Forschungsorten spezifische Fragestellungen und Datentypen behandelt. Die Datensätze, Standards und Methoden werden in den Lernorten mit didaktischen Mitteln in fallstudienbasierte und kompetenzorientierte Open Educational Resources transformiert. Assessment-Werkzeuge verbinden Forschungs- mit Lernorten auf Basis eines optimierten QUADRIGA-Datenkompetenzframeworks und ermöglichen Forschenden, ihre Datenkompetenzen entlang des Datenflusses zuzuordnen, um so die QUADRIGA-Bildungsangebote zielgruppengerecht und interessengeleitet zu nutzen. Die Fallstudien werden auf Basis von Jupyter Notebooks umgesetzt und über QUADRIGA-Space sowie QUADRIGA-Navigator zur Verfügung gestellt. Weitere Informationen: quadriga-dk.de/de/. (Kontakt: [Bettina Buchholz](mailto:Bettina.Buchholz@uni-potsdam.de), Universität Potsdam)

Gauss AI Compute Competition

Das GCS hat gemeinsam mit dem BMBF sowie den Wissenschaftsministerien der Länder Nordrhein-Westfalen und Baden-Württemberg sowie dem Freistaat Bayern den „Gauss AI Compute Competition“ ausgeschrieben. Die Durchführung des Wettbewerbs übernehmen GCS und JSC. Ziel des Aufrufs ist es, den Auf- und Ausbau des deutschen KI-Ökosystems zu unterstützen und den Einsatz von KI in der deutschen Forschung und Industrie zu fördern. Schwerpunkt sind Vorhaben, die nicht nur die Technologie effektiv nutzen, sondern auch ein hohes Maß an gesellschaftlicher Wirkung und Anbindung an die Region aufweisen. Einsendeschluss ist der 17. März. Die Gewinner-Projekte dürfen für ihre Anwendungen den künftigen Exascale-Supercomputer JUPITER des JSC nutzen. Weitere Informationen: gauss-centre.eu/news/gcs-calls-for-gauss-ai-compute-competition. (Kontakt: [Chaya Chatterjee](mailto:Chaya.Chatterjee@gcs.de), GCS)

Veranstaltungen

- 17.–21.02.2025: [Iterative Linear Solvers and Parallelization](#), HLRS@GCS
- 18.–20.02.2025: [Parallel Programming of High Performance Systems](#), LRZ@GCS
- 26.–28.02.2025: [Introduction to Parallel Programming with OpenMP](#), NHR@FAU, online
- 10.–14.03.2025: [Parallel Programming in Computational Engineering and Science 2025](#), IT Center RWTH Aachen

HPC-Kalender der Gauß-Allianz:

hpc-calendar.gauss-allianz.de