BayernKI NHR FAU lrz

AI Round Table –
State and Future of AI-based Research in Bavaria

Prof. Dr. Gerhard Wellein, NHR@FAU
Dr. Nicolay Hammer, LRZ

Presentations start at 4:30 p.m.

Bayerische Zusammenarbeit LRZ + NHR@FAU

Hand in Hand für Hochleistungsrechnen & KI in Bayern

Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

NHR FAU

FAU
Friedrich-Alexander-Universität
Erlangen-Nürnberg

HPC-/KI-Versorgung in Bayern

- **Bayernweite Grundversorgung**
  - Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften (LRZ)
  - Zentrum für Nationales Hochleistungsrechnen Erlangen (NHR@FAU)

- Projektversorgung – **Antragsbasiert** (deutschlandweit)
  - „Tier-2“: NHR@FAU – Projekte mit hohem Bedarf ☐NHR
  - „Tier-0/1“: LRZ – Projekte höchster Leistungsklasse ☐GCS

Agenda

BayernKI NHR FAU lrz

- Aktuelle Ressourcen
- BayernKI – Stand und Ausblick
- Feed-Back – KI-Nutzer NHR@FAU
- Diskussion

Alex (A100)	TOP500
#184	06/22
#174	11/22
#154	06/23
#187	11/23
#212	06/23

BayernKI NHR FAU lrz

Alex Cluster

MEGWARE AMD / NVIDIA

656 NVIDIA GPUs

352 A40 (44 nodes)

304 A100 40GB/80GB (38 nodes)

1 PB

CEPH – 7 NVMe-Storage nodes

Upcoming (Q4/24):

384 NVIDIA H100 94 GB HBM2e (96 nodes)

LRZ SuperMUC-NG Phase 2

BayernKI NHR FAU lrz

total system characteristics

- 240 compute nodes
- 480 server CPUs (AI optimised)
- 960 data center GPUs
- 123 TB DDR5 Memory

high speed interconnect

- Mellanox HDR Infiniband
- fat tree topology
- two uplinks per node
- separated from Phase 1

accelerated node architecture

- 2 Intel Sapphire Rapids CPUs
- 4 Intel Ponte Veccio GPUs
- 512 GB DDR5 main memory
- Lenovo's SD650-I v3 platform

distributed asynchronous object storage (DAOS)

- 1 PB capacity
- > 750 GB/s write bandwidth

https://doku.lrz.de/access-and-login-to-supermuc-ng-11482471.html

LRZ AI System incl. MCML Partition

System Hardware Generations 1, 2 and 3

BayernKI NHR FAU lrz

LRZ Partition (Gen 1/2, in operation)

(heterogenous partition, **90 data center GPUs**)

- 52x NVIDIA A100 (80GB HBM)
- 8x NVIDIA A100 (40GB HBM)
- 30x NVIDIA P100 / V100 (16GB HBM)
- various nodes: DGX A100, DGX-1, Lenovo, etc.
- HDR Infiniband (IB) / 100G Ethernet
- 200TB NVMe network storage (over IB)

MCML Partition (Gen 1/2, in operation)

(DGX POD / HGX POD, **148 data center GPUs**)

- 64 NVIDIA A100 (40GB HBM)
- 84 NVIDIA A100 (80GB HBM)
- 8x NVIDIA DGX A100
- 21x Lenovo SD650-N (4 GPUs)
- HDR Infiniband (IB) / 100G Ethernet
- 200TB NVMe network storage (over IB)

LRZ Partition (Gen 3, *BayernKI,* ordered)

(Lenovo HGX POD, **120 data center GPUs**)

- 120x NVIDIA H100 (92GB HBM gen2)
- 30x Lenovo SD665-N (4 GPUs)
- NDR Infiniband

MCML Partition (Gen 3, ordered)

(Lenovo HGX POD, **84 data center GPUs**)

- 84x NVIDIA H100 (92GB HBM gen2)
- 21x Lenovo SD665-N (4 GPUs)
- NDR Infiniband

https://doku.lrz.de/lrz-ai-systems-11484278.html

BayernKI NHR FAU lrz

KI-Offensive Bayern

Bayern gestaltet die revolutionären Veränderungen, die Künstliche Intelligenz (KI) in sämtlichen Bereichen der Gesellschaft mit sich bringen wird, aktiv mit. Mit über 130 neuen KI-Professuren und weiteren Maßnahmen aus der Hightech Agenda Bayern hat der Freistaat beste Voraussetzungen für ein erfolgreiches bayerisches KI-Ökosystem geschaffen. Ein kraftvolles Paket von eng miteinander verzahnten Projekten verleiht ihm jetzt zusätzlichen Schub:

1. Bayern baut eigene KI-Rechnerinfrastruktur für die Wissenschaft: Bayerns KI-Expertinnen und Experten an den Hochschulen brauchen für ihre Forschungsprojekte Zugang zu ausreichender Rechenkapazität. Daher baut der Freistaat eine Bayerische KI-Rechnerinfrastruktur für die Forschenden an den bayerischen Hochschulen auf.

Ab 2024 werden am Leibniz-Rechenzentrum in Garching (LRZ) und dem Regionalen Rechenzentrum Erlangen (RRZE) starke KI-Cluster mit Prozessoren neuester Bauart errichtet. Hierfür stellt der Freistaat im Doppelhaushalt 2024/25 im Rahmen der Hightech Agenda bis zu 55 Mio. Euro bereit.

Begleitend werden die beiden Rechenzentren ein niederschwelliges Zugangsverfahren für dieses neue Angebot entwickeln. Auch die KI-Benutzerbetreuung sowie die methodische Beratung und Unterstützung von KI-Anwendern an den Hochschulen werden das LRZ und das RRZE weiter ausbauen.

https://www.bayern.de/bericht-aus-der-kabinettssitzung-vom-6-februar-2024/

Bayern KI – state and perspectives

BayernKI

- **Basisversorgung für Forschende an bay. Hochschulen und Universitäten**

- **Betrieb, User-Support & Training** durch LRZ und NHR@FAU

- Koordination und Konzepterstellung durch **Steuergruppe** des StMWK
  - Eymann (U Bayreuth), Groß (TH Nürnberg)
  - Martin (UTN), Seidl (LMU, MCML) – Kranzlmüller (LRZ), Wellein (NHR@FAU)

- Stand:
  - Erarbeitung einer „Feinplanung“
  - Erste Hardware im Zugang zu den Zentren
  - Erste Nutzer auf Bestandshardware aktiv
  - Einrichtung eines **Nutzungsausschusses** (in Vorbereitung)

Eigenschaften – BayernKI

BayernKI NHR FAU Irz

- Einfacher, **niederschwelliger Zugang zu den BayernKI Ressourcen**
  - Schlankes & schnelles Zugangsverfahren
  - Projektlokale Verwaltung der Zugänge und Projektressourcen
  - **Nachgängige Berichtspflicht/Begutachtung** von Projekten
  - Stoßzeiten-Regelung vor großen KI-Konferenzen – Prioritäten

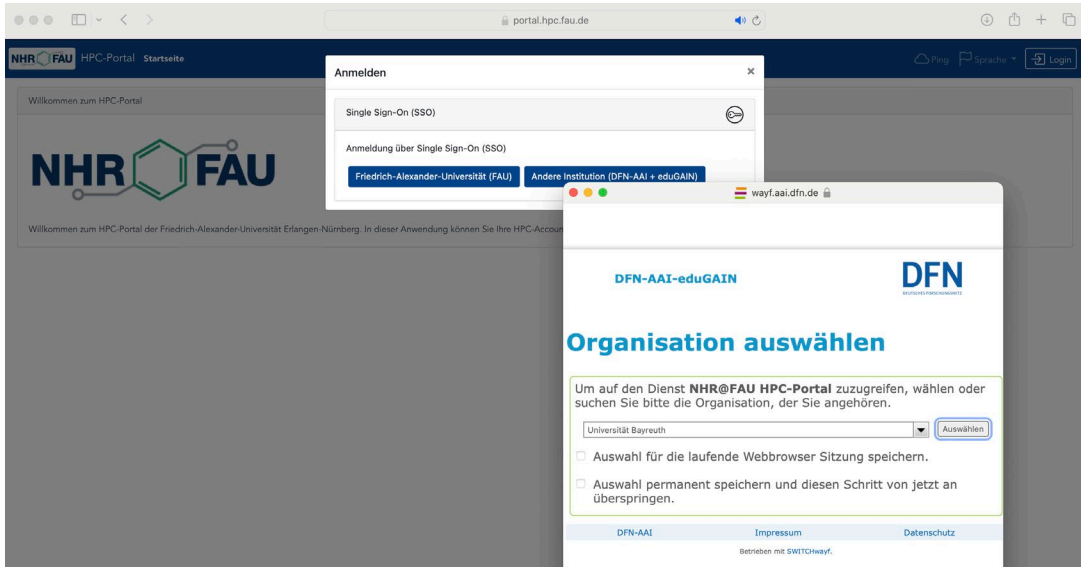
- Nutzungskonzept
  - **Nutzungsberechtigung: Forschende an bay. Univ. & HAWs**
  - Einfache techn. Zugangsverfahren/Authentifizierung (exist. Portale)
  - *Ressourcenallokation bei Überbuchung → Nutzungsausschuss / Steuerkreis*

Stand

Zugang/Projektverwaltung

BayernKI

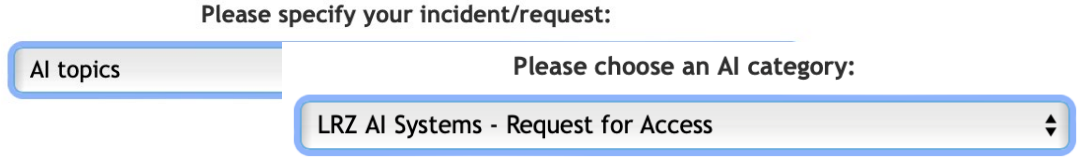
- HPC portal:
  - DFN-AAI
  - Rollen: PI – PoC – Users



Interessiert? → hpc-support@fau.de

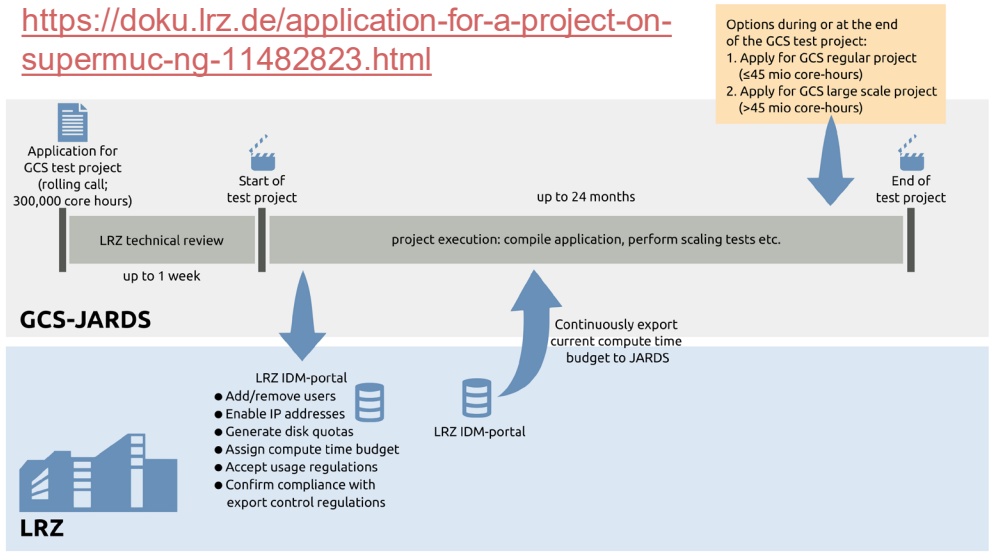
Zugang (LRZ):

- Rollen: PI – Master User – User
- Anforderung (**AI Systems**) über Service-Request (https://doku.lrz.de/lrz-ai-systems-11484278.html, https://doku.lrz.de/3-access-and-getting-started-10746642.html)



- Antrag (**SuperMUC-NG**) – GCS-JARDS Portal

https://doku.lrz.de/application-for-a-project-on-supermuc-ng-11482823.html



Interessiert? ➡ https://servicedesk.lrz.de

Eigenschaften – BayernKI

BayernKI NHR FAU lrz

- Aufbau eines **gemeinsamen KI-Betreuungs- und Schulungskonzepts**
  - Zielgruppenorientierte Trainings und Ausbildung
  - Mentoring für komplexe, ressourcenintensive Anwendungen
  - Ansprechpartner vor Ort (Universitäten/HAWs z.B. lokales RZ)
  - **Personalakquise hat begonnen – Fachberater mit KI-Expertise**

- Existierend:

  NHR@FAU: 2 Fachberater

  Das LRZ Big Data and AI Team

  Momentan 16 **direkte** or **assozierte** Mitglieder
  Team in Wachstumsphase
  Unterschiedliche **wissenschftl. Hintergründe**
  Interaktion mit anderen **LRZ-Experten**

  Unsere Mission

  Beratung, **Unterstützung** und **Training**
  Enge **Kooperation** mit ‘User-Communities’
  Innovation der Betriebsumgebung/-infrastruktur
  Suche nach Partnern

Eigenschaften – BayernKI

BayernKI NHR FAU lrz

- **Abgestimmtes Betriebskonzept** zwischen LRZ & NHR@FAU
  Installationsvoraussetzungen, Handling von Software und Daten, Überlaufkapazitäten bei Cloud-Providern, Hostingangebot

- **Gemeinsames Hardware-** und **Beschaffungskonzept**
  - Hardware im Zugang
  - Gemeinsame Ausschreibung in Vorbereitung
    inkl. Rahmenvereinbarung für bay. Univ. / HAWs

Hardwareplanungen BayernKI

BayernKI NHR FAU lrz

	LRZ	NHR@FAU
Installation Q3 – Q4/2024	**120 NVIDIA H100**	**200 NVIDIA H100**
2024 – gemeinsame Ausschreibung für bedarfsorientierte Erweiterung		
2025	Erweiterung um bis zu 880 KI-Beschleuniger	Erweiterung um bis zu 300 KI-Beschleuniger
2026		
Summe (2026)	**Bis zu 1.000 KI-Beschleuniger**	**Bis zu 500 KI-Beschleuniger**

Technologieoffen
Bedarfsorientiert
Warmwasserkühlung
Speichersysteme


lrz
Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften


NHR FAU
Erlangen National High Performance
Computing Center

BayernKI