

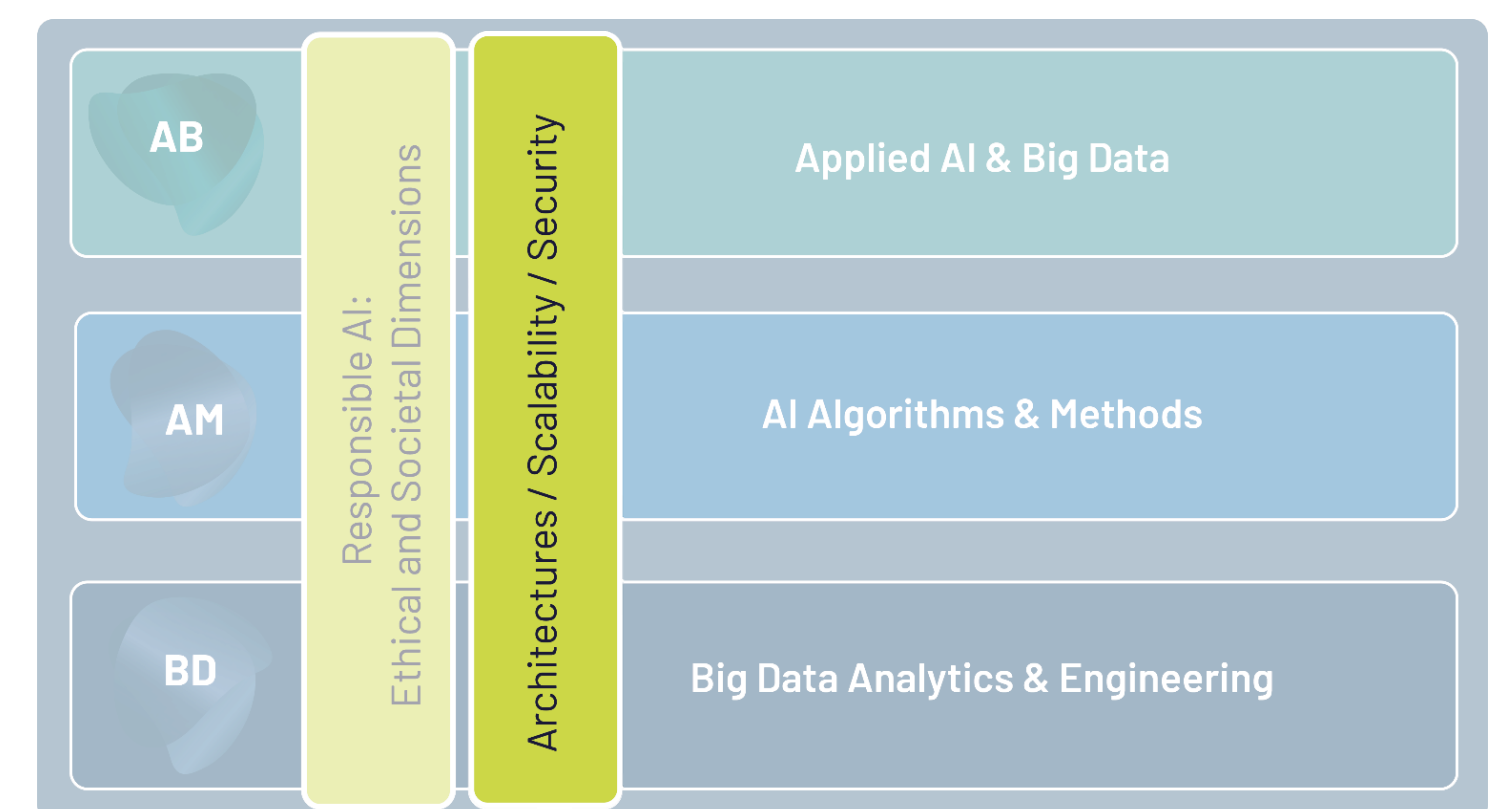
Large Language Model Services

Our Own Free Local OpenAI Alternative

Jan Frenzel^{1,2}, Dr. Taras Lazariv^{1,2}, Sima Attar², Moritz Müller^{1,2}, Apurv D. Kulkarni^{1,2}

¹ Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig

² Center for Interdisciplinary Digital Sciences (CIDS), Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH), TU Dresden, Dresden, Germany



MOTIVATION

Unlock the Full Potential of Large Language Models (LLM) with Our Local Inference Infrastructure

- Unlimited interactions without restrictions
- User-friendly GUI and flexible API access
- Privacy and data protection

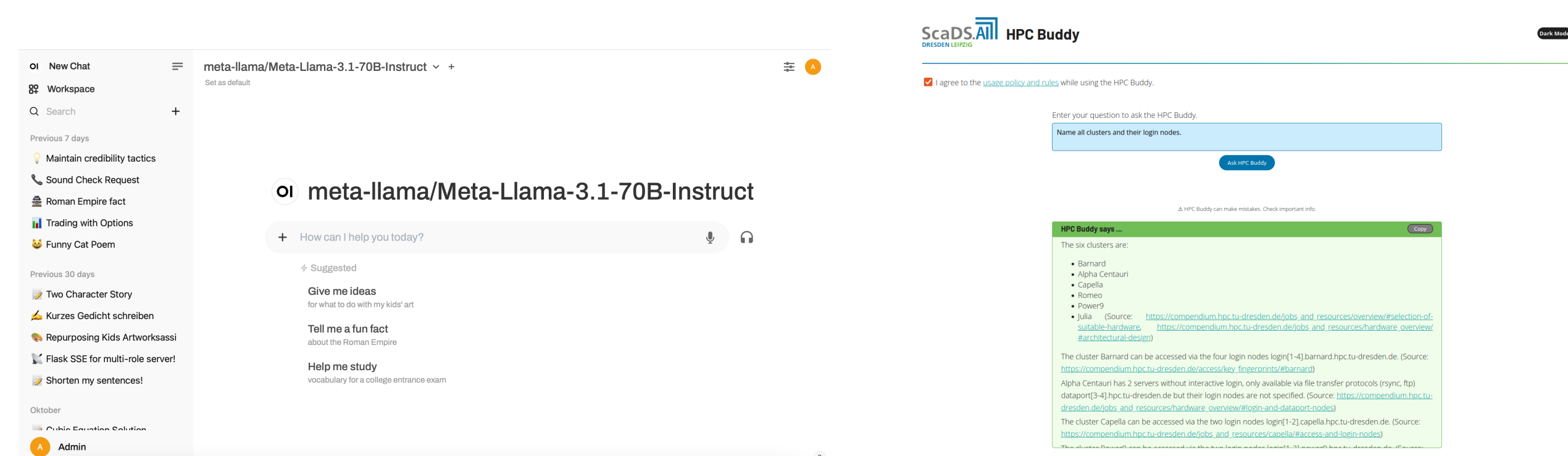
Our Mission

- Simplify access to LLMs
- Provide user-friendly services to build custom AI assistants
- Integrate LLM-based functionality into various workflows

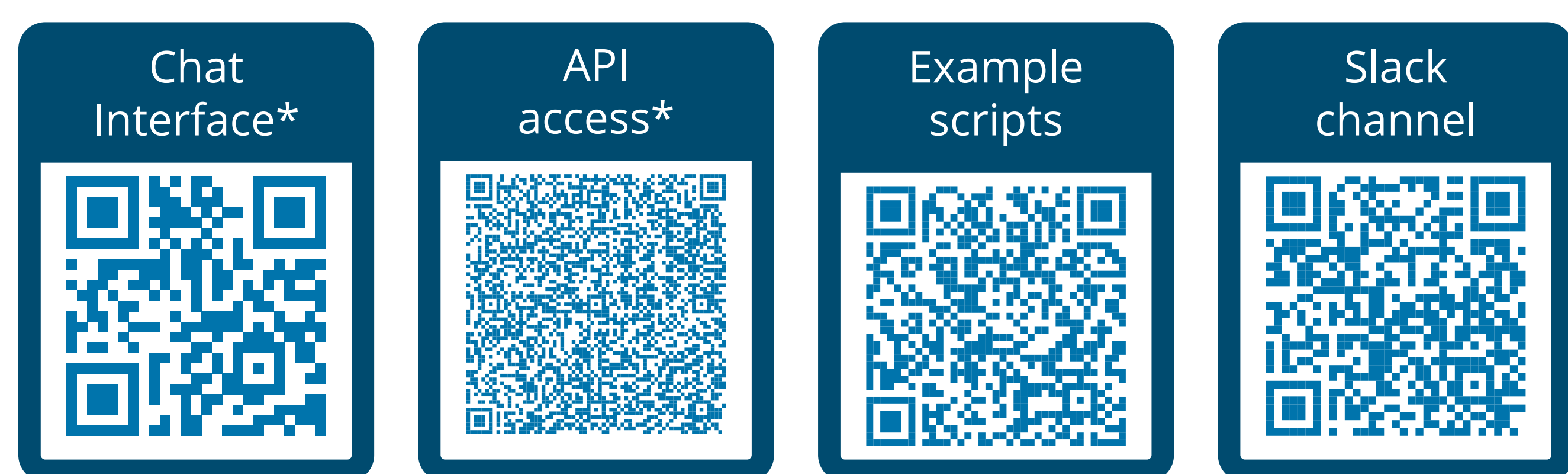
ACCESS POSSIBILITIES & INTERFACE

User-Centric Design for Easy Interaction

- API access using OpenAI-compatible API specification <https://llm.scads.ai>
- Chat interface OpenWebUI <https://chat.llm.scads.ai>
 - Use audio interaction
 - Generate images
 - Upload files and knowledge
 - Use custom tools & functions
- Slack and Matrix chat bots for integration with your group chats
- HPC Buddy - integrated custom knowledge about TUD HPC documentation



HOW TO GET ACCESS



* VPN connection to TU Dresden network required

TARGET AUDIENCE & USE-CASES

Practitioners

- Summarization: explore research papers and generate literature reviews
- Content generation: draft professional emails and assist in teaching
- Language translation: localize content and adapt communication style

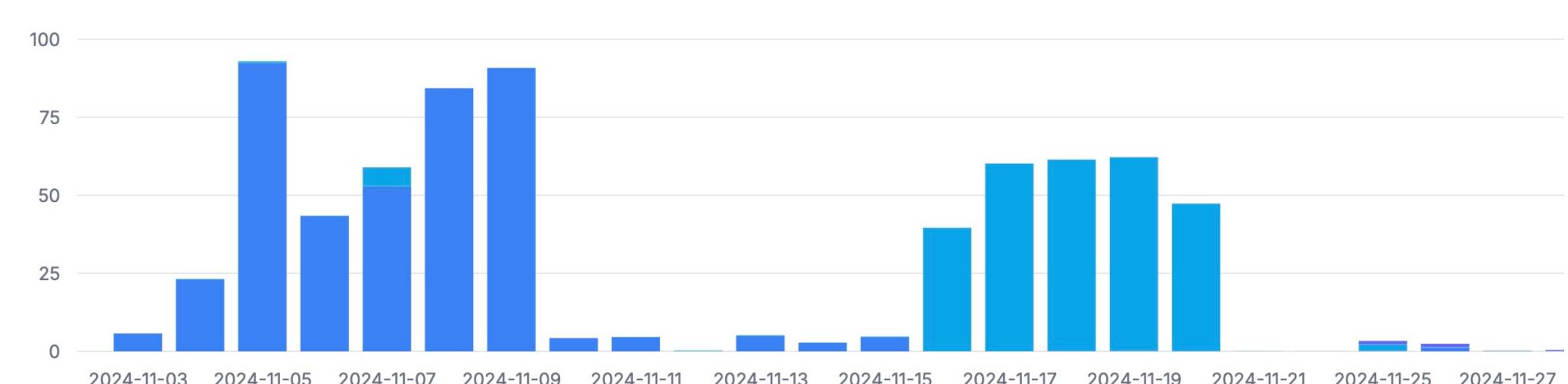
Researchers

- Natural Language Processing
- Human-Computer Interaction
- Social science research

TECHNICAL BACKGROUND

- Self-hosted open-source models on HPC nodes with focus on:
 - Multilingual conversations** Meta-Llama-3.1-70B-Instruct, c4ai-command-r
 - Coding** DeepSeek-Coder-V2-Lite-Instruct
 - Vision capabilities** Phi-3.5-vision-instruct
- Flexible Kubernetes infrastructure for all services
 - Request routing, usage monitoring, keys management
 - Custom knowledge integrations (e. g. HPC Buddy)
 - Service metrics monitoring (e. g. Dashboards)
 - Periodic jobs (e. g. knowledge ingestions)
 - Documentation

Daily Spend Per Team



OUTLOOK AND OPPORTUNITY TO CONTRIBUTE

- Retrieval Augmented Generation (RAG) services (in development): build your own expert, tailored to your knowledge
- Stay updated on the latest developments: join [#llm-inference-local-server](#)
- Share your use-case with us!
- Help develop and improve our LLM services

CONTACT

TU Dresden
Jan Frenzel
jan.frenzel@tu-dresden.de

TU Dresden
Dr. Taras Lazariv
taras.lazariv@tu-dresden.de

FUNDED BY



Bundesministerium
für Bildung
und Forschung
FKZ:ScaDS.AI



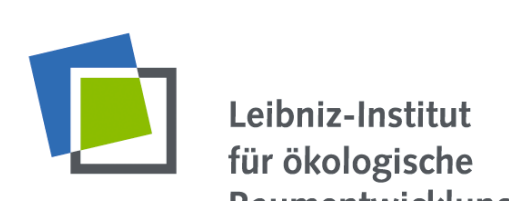
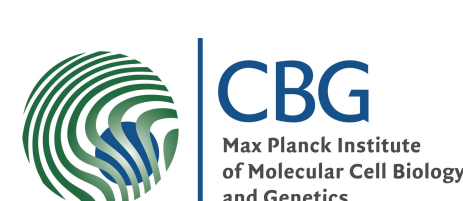
SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.



UNIVERSITÄT
LEIPZIG



MAX-PLANCK-INSTITUT
FÜR KOGNITION UND NEUROWISSENSCHAFTEN



The authors gratefully acknowledge the GWK support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden on HRSK-II.