Contribution ID: **114** Type: **Talk**

Sustainable AI Infrastructure: A Mandatory Pathway for Supercomputing Centres

Wednesday, 4 December 2024 11:40 (40 minutes)

As artificial intelligence drives more and more innovation in industry and science with the emergence of Large-Language Models (LLMs), the demand for such computing resources is growing substantially. This development poses major challenges for supercomputing centres around the world, as they need to provide state-of-the-art AI resources while meeting their sustainability goals and mandates.
While today's world-leading supercomputer according to the Top500 list require about 23 MWatt (Frontier) to provide computing power to the scientific community, we see an exponential increase in power consumption for AI training and inference. Estimates for training and interference of current LLMs are huge, and newer models will require even more power, representing an unnoticed increase in energy consumption.
For the operational requirements of HPC and AI supercomputers, we need to evaluate the total power consumption and thus the total energy based on the Power Usage Effectiveness (PUE). A four-pillar model was developed at the Leibniz Supercomputing Centre (LRZ) to implement a holistic optimisation strategy for energy efficiency, which includes the building infrastructure, system hardware and software as well as the actual applications. By extending this to AI hardware and cooling, optimising AI models and implementing AI-driven resource management, supercomputing facilities can meet these requirements while offsetting their environmental footprint.

Student or Postdoc?

PhD or DTech4

Email address

maximilian.hoeb@lrz.de

Co-Authors

CHPC User

No

CHPC Research Programme

Workshop Duration

Primary authors: HÖB, Maximilian (Leibniz Supercomputing Centre (LRZ)); Prof. KRANZLMÜLLER, Dieter (Leibniz Supercomputing Centre (LRZ))

Presenter: HÖB, Maximilian (Leibniz Supercomputing Centre (LRZ))

Session Classification: HPC Technology

Track Classification: HPC Technology