Fast Machine Learning for Science
Imperial College London
Real-time and accelerated ML for fundamental sciences
25-28 September 2023

Contribution ID: **73**

Type: **not specified**

Need for Speed: How to harness the power of Large Language Models

Wednesday 27 September 2023 11:00 (45 minutes)

Large Language Models (LLMs) will completely transform the way we interact with computers, but in order to be successful they need to be fast and highly responsive. This represents a significant challenge due to the extremely high computational requirements of running LLMs. In this talk, we look at the technology behind LLMs, its challenges, and why Groq's AI accelerator chip holds a significant advantage in running LLMs at scale.

Presenter: BECKER, Tobias (Maxeler Technologies)

Session Classification: Invited Talks

Track Classification: Invited Talks