

Machine Learning at the Edge of Scale and Speed: A Fast Future for the Next Generation of Scientific Experiments

Prof. Dr. Thea Aarrestad

ETH/ Zürich

Next-generation scientific instruments and autonomous systems increasingly rely on machine learning that must operate under extreme constraints of latency, power, and reliability. In many cases, decisions must be made in microseconds or less, often on specialized hardware such as FPGAs, ASICs, and low-power accelerators.

A compelling example is the CERN Large Hadron Collider, where real-time inference systems must process millions of events per second, rapidly filtering out irrelevant data while preserving rare signals of interest. As LHC data rates continue to grow, reaching the equivalent of 5% of global internet traffic, traditional computing approaches are no longer sufficient, requiring new methods for ultra-fast, energy-efficient machine learning.

These developments are occurring alongside emerging trends in embedded AI for wearables and IoTs, as well as highly quantized large language models and lookup-table-based inference methods designed to dramatically improve energy efficiency, reflecting a broader shift toward compact, hardware-native AI architectures.

In this presentation, we will discuss emerging techniques for low-power, low-latency inference, including hardware-aware model design, quantization, sparsity, and hardware–software co-design. Using examples from particle physics and other domains, we will show how real-time machine learning is both a practical necessity and a powerful tool for scientific discovery.