Contribution ID: **1** Type: **not specified**

Efficient training, fine-tuning and inference of large-scale ML models

Friday, 17 July 2026 11:00 (1 hour)

This talk presents model-centric methods for efficient generative AI. It explains why training and inference of LLMs are computationally heavy, then covers model compression methods such as quantization, neural network pruning, low-rank approximations, and knowledge distillation. It also introduces efficient pre-training with mixed-precision acceleration and PHEW, parameter-efficient fine-tuning methods such as LLM-Adapters, LLaMA-Adapter, P-Tuning, and LoraHub, and efficient inference techniques including speculative decoding and KV-cache optimization.

Presenter: DOVROLIS, Constantine (The Cyprus Institute)