2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) | 979-8-3503-1199-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/IPDPSW59300.2023.00008

Invited talk by Dr. Angela Dalton, Director at AMD Research

Scaling up Deep Learning: Efficiency in AI

Abstract:

In recent years, the usage of deep learning has surged. We are still in the early stages of AI proliferation, but the energy consumption for even a single training run of today's large language models is already significant. Moreover, inference is expected to be the largest energy consumer, accounting for over 70% of the total energy for a model. Growth in energy consumption for AI models is unsustainable, and unless we can find a way to drastically increase the world's energy production, we must make computing more efficient. This talk will explore the challenges of scaling deep learning in an energy-efficient way and discuss potential research directions.

Speaker Bio:

Dr. Angela Dalton is a Director at AMD Research where she leads a portfolio of research and advanced development efforts on innovative new hardware and software technologies for high performance computing, artificial intelligence, and advanced computing capabilities. She manages AMD's External Research Office, which fosters relationships with universities to drive innovation and education. Prior to joining AMD in November 2020, Angela spent twelve years at the Johns Hopkins University Applied Physics Laboratory (JHU/APL), where she was both the technical leader for strategic DOD projects and deputy director of a branch of ~250 scientists and engineers across four technical groups working to develop and advance capabilities that assure mission critical communications for US Government sponsors.

Angela has a B.S. in computer engineering from Virginia Tech as well as an M.S. and a Ph.D. in computer science from Duke University. She was an instructor and postdoctoral fellow at the University of Texas at Austin before joining JHU/APL.