Rio Yokota

Title: Scaling Laws in HPC and AI: Yesterday, Today and Tomorrow

Abstract: When computers were less powerful, and data was less abundant, many sophisticated models were invented to simulate (in HPC) or learn (in AI) the complex world around us. As the capability of computers has improved at an exponential rate for the past fifty years, brute force computing of simpler models has made sophisticated models obsolete in some areas. There are similarities and differences between how this has happened in the field of HPC and AI. For example, scaling laws in turbulence were used to bridge experiments, theory, and simulation, which are the first, second, and third paradigm of science. Scaling laws in AI on the other hand reveal the asymptotic behavior of the fourth paradigm. Scaling laws in AI comes at a time when Moore's law is approaching its end. This has many implications ranging from the design of computer architectures to the dynamics between sophisticated modeling versus brute force computing. Understanding these scaling laws is the key to predicting the dynamics between HPC and AI in the upcoming years.

Bio: Rio Yokota is a Professor at the Supercomputing Research Center, Institute of Science Tokyo. His research interests lie at the intersection of high performance computing and machine learning. He is the developer numerous libraries for fast multipole methods (ExaFMM), hierarchical low-rank algorithms (Hatrix) that scale to the full system on the largest supercomputers today. He has also lead efforts to train ImageNet in two minutes, and more recently to pre-train large language models using thousands of GPUs. He has been optimizing algorithms on GPUs since 2006, and was part of a team that received the Gordon Bell prize in 2009.