Tyler Johnson
CSE 485
4/14/2023

Seminar 2 – Advances in Accelerated Computing for AI and Scientific Computing

Speaker: Ian Buck, Vice President of Hyperscale and HPC, NVIDIA

Location: NVIDIA GTC, March 21, 2023

Ian Buck, VP of Hyperscale and HPC at NVIDIA, discusses trends in data center computing and AI in this presentation. He highlights the importance of sustainable computing, the rise of accelerated computing, and the significant impact of AI in various industries. Buck describes the need for full-stack optimization and the importance of co-designing hardware and software for specific application domains. He also explains how accelerated computing can increase energy efficiency and reduce costs. He then announces the launch of CuLitho, a library that brings accelerated computing to computational lithography, enabling faster mask creation and more masks to be created per day.

Additionally, he discusses the NVIDIA Quantum platform, which addresses the major challenges of quantum computing. He elaborates, claiming that NVIDIA Quantum addresses major algorithm and CUPU design challenges, while CuLitho empowers semiconductor fabs to achieve shorter prototype cycles, higher throughput, and reduced carbon footprint in their manufacturing process. Finally, Buck talks about pre-trained generative AI models that enterprises can customize and deploy for their applications, including content generation, summarization, sentiment analysis, translation, natural language understanding, and question answering.

Takeaways: This talk was extremely interesting to me, as the sheer scale of change in the industry of computing and energy is immense. In my opinion, we're at an inflection point of artificial intelligence, where the rate of improvement is on an exponential trajectory curve. A takeaway from this presentation that is personally applicable, is that NVIDIA has released pre-trained models, that can get smarter based on prompts fed to them. These models can be custom fit into enterprise level applications and systems through cloud based API's. Incredible stuff!