HYP_Link

Anthropic to Train/Deploy Foundation Models on AWS AI Chips

Bob Sorensen and Tom Sorensen
November 2024

RECENT DEVELOPMENT

US-based large language model (LLM) developer Anthropic and major cloud service provider (CSP) Amazon Web Services (AWS) recently announced that they were deepening their existing collaboration to support both firms' competitive prospects in the rapidly growing and increasingly competitive generative AI sector. Anthropic named AWS as its primary LLM training partner and will work with AWS to further develop AWS's AI-centric Trainum and Inferentia chips. Plans also call for Anthropic to use the AWS chips to train and deploy its future foundation models. In addition, AWS will invest $4 billion in Anthropic, adding to its earlier $4 billion investment last year, although AWS will remain a minority investor in Anthropic.

Anthropic currently offers Claude, an advanced LLM product for reasoning, vision analysis, code generation and multilingual processing, that competes against other AI-sector LLM offerings including OpenAI's GPT-4, Google's PaLM, and Meta's Llama. For its part, AWS offers an internally developed AI-centric component, the Trainium2, that reportedly can scale up to 30,000 chips capable of supporting training for 100B+ parameter models through its AWS' EC2 UltraCluster product. Deployment of the Anthropic trained models for inferencing applications will likewise be offered with the AWS Inferentia2 chip option through Bedrock, the AWS platform for hosting and fine-tuning generative models.

ANALYST COMMENT

Currently, competition within the generative AI sector is fierce, particularly in the CSP space, with a range of both AI chip and software suppliers vying for demonstrated performance gains, lower costs, and ultimately greater market shares. Microsoft has invested over $13 billion in OpenAI, another leading generative AI start-up, for the exclusive right to run OpenAI's models on Azure, its cloud-computing service. For its part, Google merged its internal AI operations with UK-based AI firm DeepMind in 2023. Similar to AWS hardware efforts, both Google and Microsoft have AI-centric chip development efforts targeted for internal use, including the Trillium TPU and Maia 100 respectively. It remains to be seen if such efforts will produce a single winner any time soon, but one thing is clear: the major CSPs are moving to foster internal AI software and chip capabilities, likely to reduce their overall dependence on NVIDIA, currently the dominant hardware/software force within the AI sector.