Ray Shines With NVIDIA AI: Anyscale Collaboration to Help Developers Build, Tune, Train and Scale Production LLMs

NVIDIA AI software integrated with Anyscale Ray unified computing framework accelerates and boosts efficiency of generative AI development with open-source and supported software.

Author: Adel El Hallak

Large language model development is about to reach supersonic speed thanks to a collaboration between NVIDIA and Anyscale.

At its annual Ray Summit developers conference, Anyscale — the company behind the fast growing open-source unified compute framework for scalable computing — announced today that it is bringing NVIDIA AI to Ray open source and the Anyscale Platform . It will also be integrated into Anyscale Endpoints, a new service announced today that makes it easy for application developers to cost-effectively embed LLMs in their applications using the most popular open source models.

These integrations can dramatically speed generative AI development and efficiency while boosting security for production AI, from proprietary LLMs to open models such as Code Llama, Falcon, Llama 2, SDXL and more.

Developers will have the flexibility to deploy open-source NVIDIA software with Ray or opt for NVIDIA AI Enterprise software running on the Anyscale Platform for a fully supported and secure production deployment.

Ray and the Anyscale Platform are widely used by developers building advanced LLMs for generative AI applications capable of powering ■intelligent chatbots, coding copilots and powerful search and summarization tools.

Generative AI applications are captivating the attention of businesses around the globe. Fine-tuning, augmenting and running LLMs requires significant investment and expertise. Together, NVIDIA and Anyscale can help reduce costs and complexity for generative AI development and deployment with a number of application integrations.

NVIDIA TensorRT-LLM , new open-source software announced last week, will support Anyscale offerings to supercharge LLM performance and efficiency to deliver cost savings. Also supported in the NVIDIA AI Enterprise software platform, Tensor-RT LLM automatically scales inference to run models in parallel over multiple GPUs, which can provide up to 8x higher performance when running on NVIDIA H100 Tensor Core GPUs, compared to prior-generation GPUs.

TensorRT-LLM automatically scales inference to run models in parallel over multiple GPUs and includes custom GPU kernels and optimizations for a wide range of popular LLM models. It also implements the new FP8 numerical format available in the NVIDIA H100 Tensor Core GPU Transformer Engine and offers an easy-to-use and customizable Python interface.

NVIDIA Triton Inference Server software supports inference across cloud, data center, edge and embedded devices on GPUs, CPUs and other processors. Its integration can enable Ray developers to boost efficiency when deploying AI models from multiple deep learning and machine learning frameworks, including TensorRT, TensorFlow, PyTorch, ONNX, OpenVINO, Python, RAPIDS XGBoost and more.

With the NVIDIA NeMo framework, Ray users will be able to easily fine-tune and customize LLMs with business data, paving the way for LLMs that understand the unique offerings of individual businesses.

NeMo is an end-to-end, cloud-native framework to build, customize and deploy generative AI models anywhere. It features training and inferencing frameworks, guardrailing toolkits, data curation tools and pretrained models, offering enterprises an easy, cost-effective and fast way to adopt generative AI.

Ray open source and the Anyscale Platform enable developers to effortlessly move from open source to deploying production AI at scale in the cloud.

The Anyscale Platform provides fully managed, enterprise-ready unified computing that makes it easy to build, deploy and manage scalable AI and Python applications using Ray, helping customers bring AI products to market faster at significantly lower cost.

Whether developers use Ray open source or the supported Anyscale Platform, Anyscale's core functionality helps them easily orchestrate LLM workloads. The NVIDIA AI integration can help developers build, train, tune and scale AI with even greater efficiency.

Ray and the Anyscale Platform run on accelerated computing from leading clouds, with the option to run on hybrid or multi-cloud computing. This helps developers easily scale up as they need more computing to power a successful LLM deployment.

The collaboration will also enable developers to begin building models on their workstations through NVIDIA AI Workbench and scale them easily across hybrid or multi-cloud accelerated computing once it's time to move to production.

NVIDIA AI integrations with Anyscale are in development and expected to be available by the end of the year.

Developers can sign up to get the latest news on this integration as well as a free 90-day evaluation of NVIDIA AI Enterprise .

To learn more, attend the Ray Summit in San Francisco this week or watch the demo video below.

See this notice regarding NVIDIA's software roadmap.

Original URL: https://blogs.nvidia.com/blog/2023/09/18/llm-anyscale-nvaie/