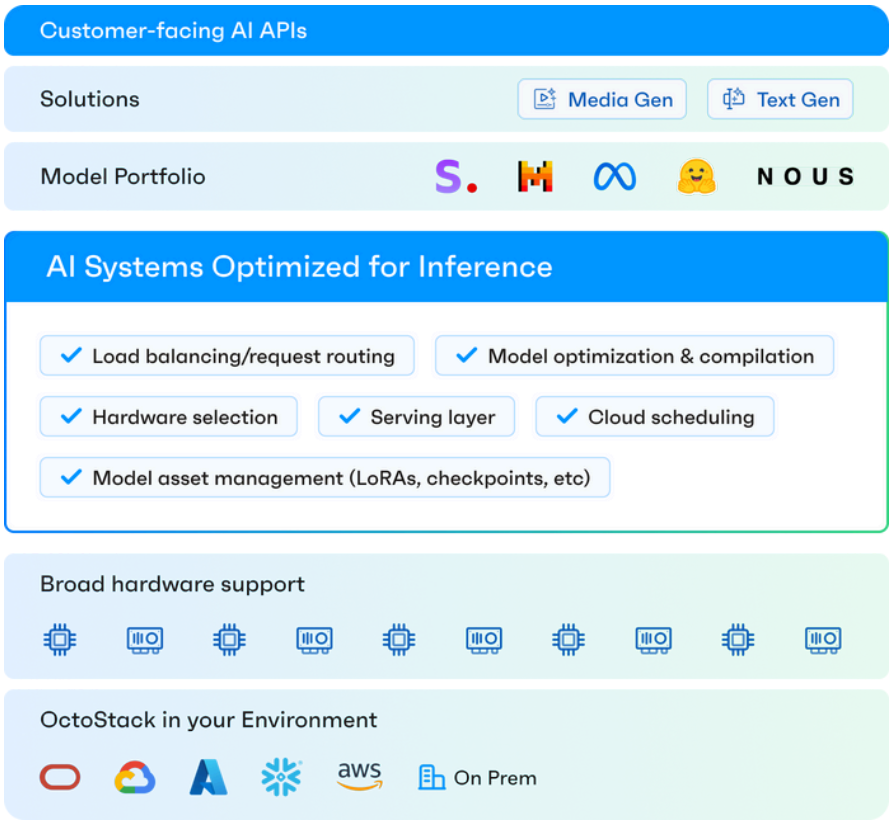
OctoAI

OctoStack: efficient, customizable, reliable GenAI inference for any environment

OctoAI is on a mission to enable customers to get value out of the latest AI innovations by offering efficient, customizable, reliable AI systems to everyone. Our leading GenAI inference platform is used by more than 25,000 developers and powers hundreds of production AI applications at massive scale. OctoAI's highly optimized AI systems stack delivers market-leading price and performance — delivering up to 80x savings versus proprietary models — without sacrificing speed or quality. OctoStack extends these capabilities to run in any customer environment for complete privacy and control over models, data, and compute.

OctoAI Systems Stack



Customizability

Build with custom or OSS models (e.g. Llama3, Mixtral, Mistral), and fine-tunes using your own data and GPUs.

Efficiency

OctoStack customers benefit from optimization at every layer of the stack for 10X faster inference vs. DIY implementations.

Reliability

Load balancing, autoscaling, and support for a wide-range of GPUs,SLAs ensure your app is supported as usage grows.

AI systems expertise built-in

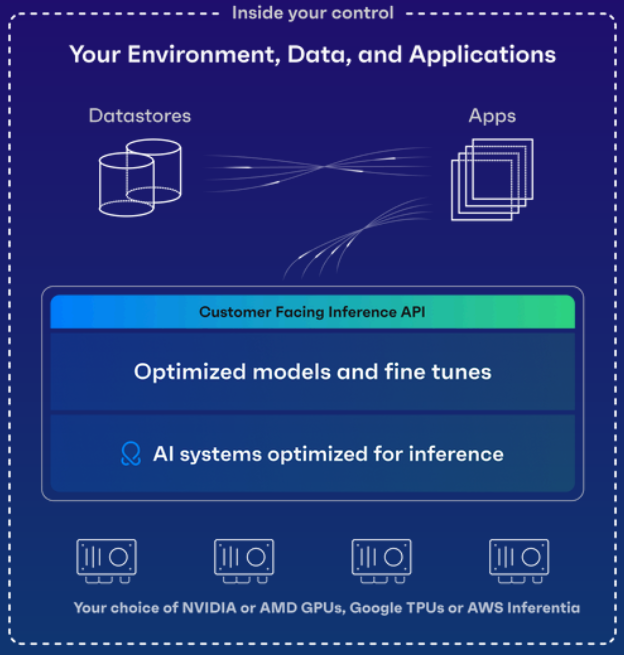
Powering OctoAI's solutions and OctoStack are world-class AI systems that automatically optimize models and improve hardware utilization with smart resource management. OctoAI's differentiated technology unlocks GPU availability, reduces cost, and improves end-user experience. OctoStack customers benefit from continuous optimization of OctoAI's SaaS service, and a subscription to newly optimized models and hardware support.

OctoAI

Your models, your data, your environment

OctoStack is a turnkey production GenAI serving stack that delivers highly-optimized inference at enterprise scale. Run and scale the latest models including Llama3, Mistral, and Mixtral from one OpenAI-compatible API endpoint. Customers can also run custom models and fine-tunes with zero extra markup. Build enterprise applications with native JSON mode, easy-to-build RAG recipes, and data service integrations (e.g. Snowflake).

> OctoStack delivers on our performance and security-sensitive use case. It lets us easily and efficiently run the customized models we need within the environments we choose and supports the scale our customers require.



AICPA SOC

SOC 2 Type II Certified

10x Performance boost vs. best-in class DIY

4x Better GPU utilization than best-in-class DIY

50% Reductions in operating costs

About OctoAI

OctoAI is home to leading experts in ML systems, model compilation, and hardware intrinsics. The company was founded in 2019 by the creators of widely adopted open source ML projects including MLC-LLM, XGBoost, and Apache TVM. Based in Seattle, WA, and backed by Madrona, Amplify Ventures, Tiger Global and Addition Capital.

OctoAI

OctoStack FAQ

What language model checkpoints can I run?

OctoStack supports the most popular base models and fine-tuned checkpoints. This includes the Llama 2 (13B and 70B), Llama3 (8B and 70B) Code Llama (7B,13B and 34B), Mistral 7B, Mixtral 8x7B, and Mixtral 8x22B. Customers can also run fine-tuned checkpoints such as Nous Hermes 2 Pro Mistral 7B, or any checkpoint created from the supported base models. Customers can easily configure the desired model weights with a single line, and model weights can be downloaded from a Hugging Face repository or cloud storage like S3.

The model catalog is constantly updated as new models are released. If there is a specific model that you need to run that is unavailable in the catalog, the OctoAI team can work with you to add it to your deployment.

What GPU hardware should we run for language models?

The best throughput and latency is typically achieved on NVIDIA A100 or NVIDIA H100 GPUs. The minimum recommended GPUs for larger models (like 70B or 8x7B) are 2 A100 or H100 GPUs. Smaller models (like the 7B and 8B models) can be run using NVIDIA A10G GPUs.

How does the deployment process work?

OctoStack supports Kubernetes deployment via a helm chart, and a Docker Compose deployment compatible with any container orchestration service. The following are the steps to deploy OctoStack in an Amazon VPC:

- OctoAI provides access to your AWS account, so you can retrieve OctoStack containers in OctoAI's AWS ECR container registry
- You pull the OctoStack containers from OctoAI's AWS ECR
- You configure and deploy using Kubernetes or Docker Compose
- We provide a simple guide and example configuration files. We also provide a helm chart to make Kubernetes deployments easier

How does load balancing work in OctoStack?

OctoStack includes an advanced load balancing solution to increase GPU throughput and utilization. We've developed a load balancing system that is designed for generative AI workloads, and can manage both homogenous and heterogenous workload profiles.

How does auto-scaling work in OctoStack?

OctoStack emits metrics allowing you to decide when to scale GPU's up and down. Emitted metrics include the number of pending and in-flight requests, and requests per second.

What options are available for using embeddings or building retrieval augmented generation (RAG) workflows?

We have integrations and partnerships with industry leading services, including Langchain, Unstructured, LlamaIndex, Pinecone, and others. You can find details of these integrations and examples in the OctoAI docs pages. All OctoAI integrations are available in OctoStack.