Lenovo
AI Innovators
Partner Ecosystem

EDGE BRICKS

Lenovo and NVIDIA power Edgebricks' ML Platform as a Service Solution for AI Models

Edgebricks' platform provides a fully Automated Machine Learning Infrastructure to build, train, and deploy AI Models 10 times faster

---

Summary

89% of enterprises today are looking to do more with AI and Machine Learning adoption in their existing products, processes, and day-to-day work. Generative AI is causing a further increase in potential applications by 100x as we go from classification-based AI to generating content for images, videos, code, text and other media types.

Building and running private GPT models is essential for many industries due to compliance, data privacy and cost reasons. Machine Learning is also the most expensive workload to host in a public cloud since it requires high compute, high memory, network and storage bandwidth.

Edgebricks AI/ML platform as a service helps enterprises build, train and deploy AI/ML 10x faster using Lenovo's Edge servers with NVIDIA GPUs at the location of their choice. So they can get ease of use of a public cloud but with control and flexibility of a private AI/ML cloud.

Challenge

The growth of Enterprise AI and the adoption of Generative AI are causing companies to adopt AI/ML to remain competitive, increase their team's productivity and lower costs. They either have to build the AI/ML infrastructure by themselves or leverage some one-size-fits-all solution on public clouds. However, most of them mention hiring the right talent for AI/ML as one of their biggest problems.

The overall set of tools and landscape is evolving fast and there is no easy way to build AI/ML infrastructure and pipelines without high cost and expertise. Public clouds offer ready made solutions but their cost increases significantly as adoption increases.

Furthermore, training an AI model is also a time consuming process and companies interested in having tailored Generative AI models are pressed to launch the solution in a short time frame to improve their performance and free assets for more important tasks.

Lenovo | NVIDIA

Solution

Using Lenovo's Edge servers powered by NVIDIA GPUs, Edgebricks offers a solution that allows companies to build, train and deploy AI/ML models 10 times faster at the location of their choice. This gives companies ease of use of a public cloud but with control and flexibility of a private AI/ML cloud.

Edgebricks AI/ML PaaS solution is comprehensive and end-to-end allowing companies to simply start by running their pipelines. The platform grants access to your AI/ML engineers allowing them operate in a self-service manner.

We provide a unique model where you can offload expensive stages of the pipeline that do the training to on-premises or Colo infrastructure. We provide a built-in object store in Edgebricks cloud to store this data and run training jobs on the servers with fast access to data on the same LAN. There is no need to pay high data storage or access costs in the cloud.

Results

Using Edgebricks platform as a service, enterprises can build AI/ML pipeline within days. Edgebricks supports the following AI projects out of the box: object detection, NLP, recommendation systems, and private LLMs.

Once the infrastructure is in place, Edgebricks can convert that into AI/ML cloud within a day and help build, train, test, and deploy AI models without enterprises needing a lot of DevOps expertise. Developers get a self service portal to test and deploy applications. DevOps teams get automated pipeline building workflows, so they don't have to stitch everything together themselves.

The three key benefits include:

- Build, train, and deploy models 10x faster as compared to doing it yourself
- Get best-of-the-breed apps with automation to deploy them
- Lower overall TCO by 80% as compared to using public clouds

In addition to these benefits, customers also get data privacy and compliance as they can deploy this infrastructure behind their own firewall at the location of their choice.

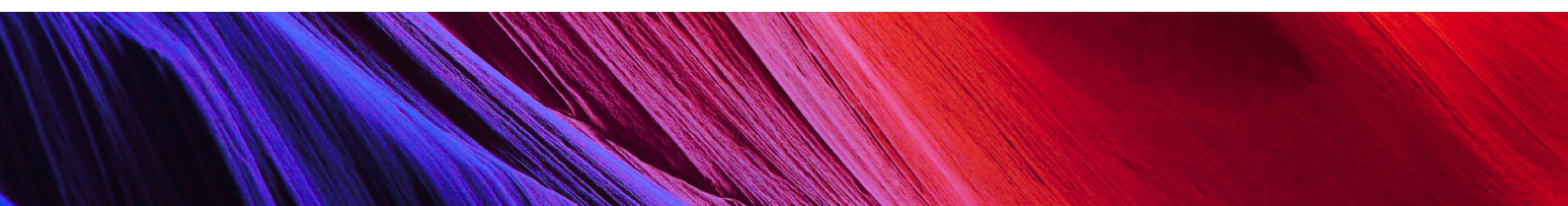
Additional benefits include:

- Setup AI/ML infrastructure and pipeline in few hours
- Lower total cost of ownership by 80%
- Provide self-service to AI dev team
- Provide control, visibility and low opex to devops team

Validated Architecture

This solution uses Lenovo ThinkEdge SE450 and SR650 V2 servers as the core infrastructure layer. These servers will have Intel Xeon CPU with anywhere between 16 to 32 cores per CPU, up to 1TB of RAM, SSDs for fast storage, and NVIDIA A100 Tensor Core GPUs for the most demanding AI/ML workloads. Customers will also get the best-in-class support and warranty from Lenovo.

Edgebricks edge software will convert these servers into an AI/ML cloud. Customers can then consume this cloud using a simple web portal. Edgebricks provides a built-in app store with AI/ML tools and a pipeline builder to accelerate AI projects in an enterprise.



There is a built-in App Store with dozens of AI/ML tools. Deployment and integration of these tools is automated, to build various workflows and pipelines as needed. These include automation of object detection, data labeling, NLP, recommendation systems, and others.

Design Components

Servers	Other Specs	Accelerator	Use Case	Software
ThinkEdge SE450	multi-gig networking and local SSD per host	NVIDIA A30	AI/ML Build & Inference	Edgebricks AI/ML PaaS
ThinkEdge SR650V2	multi-gig networking and local SSD per host	NVIDIA A100	AI/ML Worklaods at Scale	Edgebricks AI/ML PaaS
ThinkEdge SE450	multi-gig networking and local SSD per host	-	Edge / Colo Private Cloud	Edgebricks Edge/Cloud Infrastructure

Resources

- Explore Lenovo's AI Innovators Program
- Explore the Lenovo HPC and AI Innovation and Briefing Center
- Lenovo Validated Design for AI Infrastructure on ThinkSystem Servers
- Edgebricks Website
- Edgebricks - Build AI/ML Pipelines in Minutes
- Lenovo-NVIDIA Alliance

Why Lenovo

Focused on a bold vision to deliver smarter technology for all, Lenovo is developing world-changing technologies that create a more inclusive, trustworthy, and sustainable digital society. By designing, engineering and building the world's most complete portfolio of smart devices and infrastructure, we are also leading an Intelligent Transformation to create better experiences and opportunities for millions of customers around the world.

Why NVIDIA

NVIDIA pioneered accelerated computing to tackle challenges no one else can solve. Our work in AI and the metaverse is profoundly impacting society and transforming the world's largest industries—from gaming to robotics, self-driving cars to life-saving healthcare, climate change to virtual worlds where we can all connect and create.