Rokid Scales AI Infrastructure with [SMART]Rack AI Deep Learning Platform

OVERVIEW

Rokid is an AI company focused on the development of voice AI-enabled products, such as smart home assistance technology, smart speakers, and companion robots capable of empathy. For their next deployment phase, they wanted a partner who had a strong background in designing high-performance computing architecture, as well as expertise in AI/Deep Learning frameworks to ensure their partner could speak the same language as the key users of the solution.

Because AMAX Deep Learning platforms have a strong reputation in the industry, particularly for being the brand of choice for ImageNet large Scale Visual Recognition Challenge (ILSVRC) winners, AMAX was the partner they tasked with designing and building a cluster that would perform optimally for their specific workloads.

SUMMARY

- Rokid is a Smart Home AI startup with a focus on voice-enabled products.
- Rokid needed a partner to build a new in-house GPU cluster for Deep Learning model training that fulfilled compute requirements and could easily scale.
- AMAX customized a design based on its [SMART]Rack AI platform, ensuring optimal performance and manageability that could scale at the rack level.

Rokid™

Company: Rokid™
Location: San Carlos, Silicon Valley
Industry: Smart Home AI
AMAX Products Utilized: [SMART]Rack AI, [SMART]DC

THE PROBLEM

Rokid was looking to build a new type of rackscale GPU-powered DL infrastructure to ramp up their AI development. The goal was to set up a rack featuring 8 nodes of high-density GPU servers to achieve 64 cards in total. They also had a short time frame in which they needed the cluster delivered, as they were in the midst of intense development cycles and did not want to disrupt any progress.

"Prior to this expansion, we were mostly using CPU-only systems to develop our AI algorithms," said Frank Rao, Director of Rokid Research Lab. "We found these to be highly inefficient. We needed a much more powerful rackscale compute platform to accelerate our time to market against a very aggressive development schedule." During the planning process, they also became aware of three major challenges. The first challenge was to meet the AC power requirements for such high power-density deployment. The second challenge was to determine which GPU card within NVIDIA's lineup would provide the best performance and value. Furthermore, due to a mission critical development schedule, they not only needed a solution quickly, but it was mandatory that the solution be production ready and could work out of the box.

> "Our requirement for computational power, power density as well as network speed are very different from conventional servers. AMAX was able to deliver a rackscale GPU-based solution with comprehensive power management, ultra-fast network speed and, what is even better, a fully customized Deep Learning solution that solved our problem above and beyond what we thought we were looking for."
>
> - ***Frank Rao,*** *Director of Rokid Research Lab*

THE SOLUTION

After talking to several of NVIDIA's Elite Partners, Rokid chose AMAX to build their infrastructure. The decision was based on the fact that AMAX has a strong reputation in the industry for building high-performance clusters, an engineering team experienced in optimizing infrastructure for Deep Learning, and a ready reference architecture for a turnkey Deep Learning cluster, the [SMART]Rack AI. Furthermore, AMAX had a New Product Introduction (NPI) Program that could develop, test and validate a solution in a short amount of time to ensure optimal field performance and stability.

Rokid, like many other software-centric companies, lacked experience in deploying and scaling large-scale on-premise infrastructure. To kick off the design phase, AMAX's engineering team worked closely with Rokid to map out all of their requirements, creating a long-term blueprint for the solution and the deployment plan.

The solution AMAX produced was a customization of the [SMART]Rack AI, featuring 8x DL-E280 Deep Learning servers for a total of 64 Titan X cards, providing 704TFLOPs of compute within a single rack. The solution integrated the latest 25GbE high-performance networking for increased in-rack bandwidth and productivity, removing bottlenecks between compute and storage to accelerate applications. To address Rokid's previous encounters with disruptions caused by small power glitches and brownouts (intermittent drops in voltage), AMAX integrated an in-rack battery that provides 2.5min of backup power at 10kW to safely shut down servers without requiring an external UPS.

Furthermore, since temperature/power control often presents productivity issues to single-room HPC deployments, AMAX integrated its [SMART]DC Data Center Manager to the top of the rack to enable policy-based monitoring and data analytics. [SMART]DC differentiates itself from other DCIM solutions in its ability to monitor and manage GPU-integrated hardware, along with server platforms across various manufacturers in a heterogeneous infrastructure. [SMART]DC gave Rokid insight into temperature fluctuations that were causing systems to shut down, as well as the ability to control power efficiency to optimize overall infrastructure performance during peak and idle times.

"Initially installed for remote server and in-rack battery management, we quickly found that [SMART]DC is a key asset for GPU deployments," said Rao. "When [SMART]DC reported anomalous GPU server inlet temperatures, temperatures were so high we thought there was a problem with the temperature sensors. It allowed us to discover we had a severe issue with the room's cooling system. To protect our investment, we are now using [SMART]DC's intelligent policies feature to automatically shut down equipment in times of insufficient cooling."

Together with an in-rack cooling system located at the back of the rack, [SMART]DC's advanced analytics and policy automations enabled Rokid to achieve the best resource and power efficiency.

"Our requirement for computational power, power density as well as network speed are very different from conventional servers," said Rao. "AMAX was able to deliver a rackscale GPU-based solution with comprehensive power management, ultra-fast network speed and, what is even better, a fully customized Deep Learning solution that solved our problem above and beyond what we thought we were looking for."

With Rokid's version of the [SMART]Rack AI now deployed, Rokid is seeing drastic acceleration of training cycles required to support the exponential growth of the company, while having reduced the overall cost of operations by 50%.

AMAX Information Technologies | 1565 Reliance Way, Fremont, CA 94539 | 1.800.800.6328 | www.amax.com