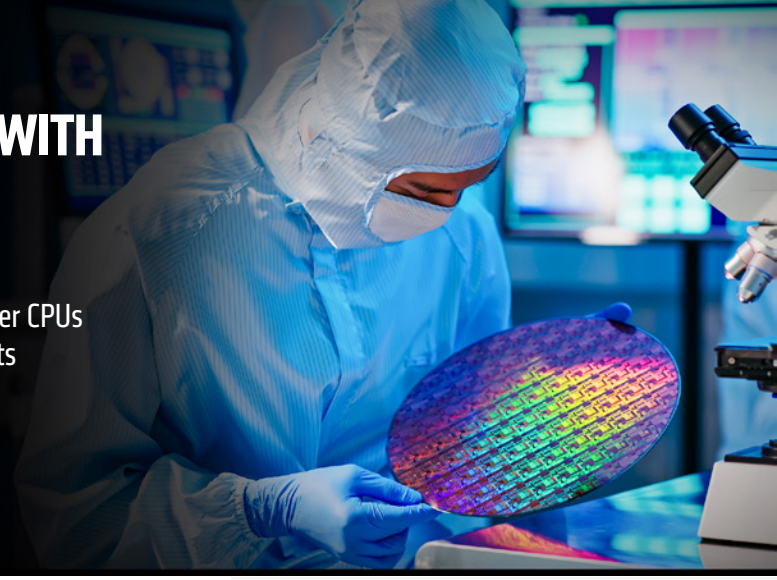
AIBIZ BOOSTS SEMICONDUCTOR YIELD WITH LEAN AI POWERED BY AMD

CASE STUDY

AIBIZ enables efficient CPU-based AI inference with AMD EPYC™ Server CPUs to detect wafer defects early, boost semiconductor yield, and cut costs



AMD × AIBIZ

AIBIZ deploys AI to catch wafer defects before they cause failures. AMD EPYC™ Server CPUs provided the performance the company needed to achieve this in a lightweight form without GPU acceleration.

"We are a six-year-old startup," says Seung-Jae Ha, CEO of AIBIZ. "Our product is called DutchBoy." This name comes from a fable about a young lad living in the Netherlands who blocks a hole in a dam with his finger, preventing a bigger disaster. The AIBIZ software performs a similar feat for semiconductor wafers.

> **"We can find small issues at the beginning of the fabrication process to increase the yield of the foundries."**
> Hyun Jin Choi, CTO, AIBIZ

"We can find small issues at the beginning of the fabrication process to increase the yield of the foundries," says Hyun Jin Choi, CTO, AIBIZ. "A foundry consists of a lot of different manufacturing processes. The quality check is usually done at the very end of production, so the defects are detected at a very late stage. But by using sensor data, we can identify problems as they occur, so they can be tackled earlier."

This is not an easy task, however. "The server that we work with is installed within the Samsung Electronics plant inside the etching equipment itself," says Choi. "We collect real-time sensor readings from 20 pieces of equipment, at a sampling rate of 100 milliseconds from 300 sensors in each, so we end up with a vast amount of data."

AMD EPYC SERVER CPUS DELIVER MULTITHREADING STRENGTH

DutchBoy works through a Docker container-based system. "We connect 20 of our products to one server," says Choi. "We use a Time Series Anomaly Detection AI analysis model and a graph neural network (GNN)." High-energy electromagnetic fields within the foundry chambers create plasma to etch the wafer. "Inside the chambers, there are sensors detecting the temperature and pressure. Unlike other AI workloads, our GNN analyzes sensor correlations and enables consolidation."

This consolidation allows AIBIZ to reduce the number of parameters in its models. "Very large parameter quantities necessitate the use of GPUs," says Choi. "GPUs typically generate more heat and require additional cooling, which can create challenges inside a fab." There are also latency issues between the GPU and CPU. "But because we're able to consolidate parameters, we can maintain our model at a very small size. This allows us to avoid using a GPU and instead use CPUs."

INDUSTRY
Semiconductor fabrication

CHALLENGES
AIBIZ faced CPU bottlenecks analyzing massive real-time sensor data in fabs, needing faster multithreading performance without heat and latency from GPUs

SOLUTION
AIBIZ deployed AMD EPYC™ 9355 and 9554 Server CPUs to run compact AI models on CPUs without GPU acceleration

RESULTS
AMD EPYC CPUs delivered 30% faster AI inference, removed bottlenecks, and enabled AIBIZ to detect wafer defects early and improve semiconductor yield

AMD TECHNOLOGY AT A GLANCE
AMD EPYC 9554 CPUs
AMD EPYC 9355 CPUs

TECHNOLOGY PARTNER
HPE

A typical current LLM might have up to 120 billion parameters, but AIBIZ's models use fewer than 100,000 parameters. While GPUs are still needed for training, the inference speed required is within the capabilities of a fast enough CPU. "Although we create the sampling data every 0.1 second, we don't need the inference that quickly," says Choi. "For the data that is generated in the 20 units that we have, from collection, to pre-treatment, to inference, to visualization, to reporting, five minutes is fast enough."

AMD EPYC Server CPUs enable AIBIZ to achieve 30% faster inference.

AIBIZ explored AMD EPYC Server CPUs after finding its existing processors couldn't deliver the performance required. "We were using Intel before, but we kept experiencing bottleneck problems," says Choi. "We realized that the way to deal with this bottleneck issue was multithreading. We were looking for a processor with strong multithreading, which led us to AMD."

AMD ENABLES SUBSTANTIAL PRODUCTION SAVINGS

AMD provided AIBIZ with initial test servers, but hardware partner HPE also provided crucial assistance. "HPE helped us identify which EPYC product would be the best," says Choi. "Our product can go not only inside the Korean plant but could also expand to other geographies like China and America, where Samsung has factories. HPE's global support was one of the reasons why we wanted to work with the company."

> **"We were looking for a processor with strong multithreading, which led us to AMD."**
> Hyun Jin Choi, CTO, AIBIZ

To test performance, AIBIZ benchmarked inference with different CPUs. "We conducted comparisons on five domain datasets with our previous Intel CPUs," says Choi. "We saw a 30% AI inference performance increase with AMD EPYC Server CPUs. After optimization, we were able to completely solve the bottleneck issue."

This has enabled AIBIZ to optimize Samsung's semiconductor production. "DutchBoy was able to provide better predictions," adds Choi. "There is a phenomenon called arcing, where you see lightning happening inside the chamber. This affects the wafer and will lead to a defect. In those arcing instances, we saw a data spike. When this occurs, we provide real-time information to the engineers so that they can prevent the wafer defect from happening."

> **"We saw a 30% AI inference performance increase with AMD EPYC Server CPUs."**
> Hyun Jin Choi, CTO, AIBIZ

These early predictions can have very real benefits for the bottom line of silicon wafer manufacture. "Memory wafers are very expensive," says Choi. "Each wafer costs about 20 million won ($13,500). Before our technology, companies would sample one from every 20 wafers." A defect could mean rejecting the entire set. "If DutchBoy finds a problem before it happens, they can save 20 wafers. This can lead to 380 million won ($255,000) of savings. Because the sampling used to be done once every day, the savings would be multiplied by 365 days meaning 150 billion won ($100 million) saved per year."

LIGHTWEIGHT AI POWERED BY AMD EPYC SERVER CPUS

"That is just one part of the process," adds Choi. "DutchBoy allows a yield improvement of 3% to 5%, and in a fab, that will equal trillions of won (billions of dollars) savings." AIBIZ's technology also prevents delays in supply. The lead time for fabrication can be up to eight months, and in an age of heavy demand, that's a game-changer for the industry.

GPUs are not required thanks to the performance of AMD EPYC Server CPUs.

"We are currently using AMD EPYC 9355 and 9554 Server CPUs at Samsung and Xian (China) Plants," says Choi. "We're also going to use the same processors for Outsourced Semiconductor Assembly and Test (OSAT) facilities and LG Innotek."

AMD will remain central to AIBIZ's technology for the foreseeable future. "We are going to expand within the foundries industry, to Hynix, Intel, and Micron, using Samsung as a reference," says Choi. "We're also going to expand within the industry ecosystem. Although many of these companies are Korean, they operate globally, so we will be expanding globally with them. We are looking forward to having AMD EPYC Server CPUs as a part of our global expansion."

> **"We are looking forward to having AMD EPYC Server CPUs as a part of our global expansion."**
> Hyun Jin Choi, CTO, AIBIZ

AMD EPYC Server CPUs enable a more lightweight approach to AI inference. "The biggest challenge with AI for our customers is always the price of the hardware, but because we don't have to rely on GPUs anymore, we can take away a lot of the cost," says Choi. "Most startups started out thinking that AI is GPU. But now we are convinced that AI does not equal GPU, and it does not compromise performance. Other companies should test AMD EPYC Server CPUs and rid themselves of that stereotype."

"Our lightweight AI process is possible only when inference is done on the CPU," says Ha. "If the system is a CPU plus GPU system, then it will be more expensive, consume more power, and produce more heat. The combination of DutchBoy and AMD is what allows a lightweight AI system to be created."

AMD EPYC Server CPUs empower AIBIZ DutchBOY lightweight AI.

WANT TO LEARN HOW AMD EPYC PROCESSORS MIGHT WORK FOR YOU?
Sign up to receive our data center content:
amd.com/epycsignup

ABOUT AIBIZ

AIBIZ is a South Korean industrial AI company focused on transforming manufacturing through process diagnostics and real-time defect analysis. Its flagship platform, DutchBoy, uses sensor data and deep learning to detect anomalies, classify defect patterns, trace affected equipment and processes, and support root-cause analysis. Serving sectors such as semiconductors and smart factories, AIBIZ helps manufacturers improve yield, reduce losses, and make faster, data-driven quality decisions. It also supports scalable adoption across sites. For more information visit ai-biz.net.

ABOUT AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics, and visualization technologies. Billions of people, leading Fortune 500 businesses, and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible. For more information about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn, and X pages.

DISCLAIMERS

All performance and cost savings claims are provided by AIBIZ and have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein are specific to AIBIZ and may not be typical. GD-181

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.

COPYRIGHT NOTICE