



FOR IMMEDIATE RELEASE

Fixstars and Macnica Offer End-to-End Support for Tenstorrent AI Products, from Hands-on Evaluation to Model Optimization, Deployment, and Operation

Joint PoC Cuts Per-Token Generation Time of OpenAI's "gpt-oss-120b" by 31.3%

Oct 02, 2026 – Fixstars Corporation (TSE Prime: 3687, US Headquarters: Irvine, CA), a leading company in performance engineering technology, and Macnica, Inc. (Headquarters: Yokohama, Kanagawa, Japan; President and Co-CEO: Kazumasa Hara) will jointly support companies in Japan that are considering the adoption of Tenstorrent AI products, covering everything from hands-on evaluation of AI models on actual hardware to performance optimization, AI infrastructure construction, deployment, and operation.



In addition, to verify the effectiveness of the technical support the two companies provide together, they conducted a joint proof of concept (PoC) to analyze and optimize the inference performance of OpenAI's open-weight large language model "gpt-oss-120b" on Tenstorrent's AI workstation "TT-QuietBox." As a result, they accelerated the MoE (Mixture of Experts) expert matrix multiplication, the main bottleneck in the decode phase, by 3.68x and reduced TPOT (time per output token after the response begins) by 31.3%.

Through this collaboration, companies can validate the AI models they plan to deploy on new AI hardware, using the actual hardware and without committing to a purchase. Based on the

results of this validation and performance optimization, they can quantitatively determine whether the AI platform suits their own use cases and performance requirements.

Background of the Collaboration

As enterprise use of generative AI expands, there is growing demand to run AI models in environments under a company's own control, for reasons such as confidentiality, governance, and operational flexibility. However, adopting new AI hardware requires more than product selection. Companies must also confirm that the AI models they actually intend to use run properly, verify performance on real hardware, tune software, build AI infrastructure, and secure technical support after deployment.

Support Provided by the Two Companies

To address these needs, the two companies will combine Macnica's support for AI infrastructure construction and product evaluation with Fixstars' performance engineering to jointly provide support spanning hands-on evaluation, performance optimization, deployment, and operation.

<Key Support Provided by the Two Companies>

- Hands-on evaluation of the AI models and applications customers use
- Construction of evaluation and model execution environments
- Porting, functional verification, and performance analysis of AI models
- Identification of processing bottlenecks
- Improvement of software implementation, including computation and data transfer methods
- Optimization of AI model inference performance
- Support for AI infrastructure construction, deployment, and operation
- Technical collaboration with Tenstorrent

This enables customers to compare and validate AI platforms based on their use cases and performance requirements, review the results of performance optimization, and then consider deploying the environment best suited to their needs.

Overview and Results of the Joint PoC

The two companies conducted a joint PoC to improve the inference performance of OpenAI's open-weight model "gpt-oss-120b" on Tenstorrent's AI workstation "TT-QuietBox" (eight Wormhole chips, 96 GB of GDDR6 memory in total).

Macnica provided the TT-QuietBox, a remote evaluation environment, the AI model execution environment, and technical support, while Fixstars conducted a detailed analysis of how the

AI model was being processed. Based on that analysis, Fixstars identified the most time-consuming parts of the processing and improved the computation and the way data is transferred.

As a result, the MoE expert matrix multiplication, which had been the main bottleneck in the decode phase, was accelerated by 3.68x. This reduced the overall TPOT of the model by 31.3%, from 144.85 ms to 99.49 ms.

In addition, per-user decode speed improved by 45.7%, from 6.9 tok/s to 10.05 tok/s, and the average end-to-end latency per request (E2EL) was reduced by 23.0%, from 25.0 seconds to 19.3 seconds.

Through this PoC, the two companies have built up technical know-how for analyzing and optimizing large-scale AI models on Tenstorrent products. Detailed evaluation conditions, analysis methods, optimizations, and measurement results are presented in a technical report published by Fixstars.

Fixstars Performance Engineering Series 06:

[Accelerating gpt-oss-120b Inference on Tenstorrent TT-QuietBox in Practice](#)

Value for Customers and Future Plans

This collaboration combines Macnica's support for AI infrastructure construction and product evaluation with Fixstars' performance engineering. Rather than simply supplying AI hardware, the two companies run the AI models that customers actually want to use on real hardware and analyze their processing in detail. Based on the analysis, they then improve the software implementation to help customers make more effective use of the AI hardware's performance. Going forward, the two companies will leverage the insights gained from this PoC to extend their support beyond gpt-oss-120b to the wide range of AI models and workloads that customers wish to use. By supporting everything from AI platform selection to performance optimization, deployment, and operation, they will contribute to further advancing the use of AI by enterprises.

Comments from the Two Companies

Representatives of the two companies commented on the collaboration and the PoC as follows.

Fixstars Corporation

Satoshi Miki, President & CEO

Performance engineering, Fixstars' core strength, is the technology of drawing out the maximum performance of AI models in line with the characteristics of the hardware. If companies can raise productivity more efficiently with computing resources of the same scale

and performance, they can hold down capital and operating costs and further increase the return on their AI investments.

Through our collaboration with Macnica and Tenstorrent, we will help customers select and make full use of their AI platforms.

Macnica, Inc.

Hideomi Ozaki, Vice President, Altima Company

As the choices of AI models and AI hardware expand, companies need to identify the combination best suited to their use cases and performance requirements. Macnica will connect Fixstars' software optimization technology with Tenstorrent's open AI platform to support product selection, hands-on evaluation, performance improvement, deployment, and operation. Through this three-company collaboration, we will help create an environment in which customers can adopt new AI technologies with confidence.

###

About Macnica, Inc.

Macnica is a service/solution company that handles the latest technologies in a comprehensive manner, centered on semiconductors and cyber security. Developing business in 100 locations in 33 countries/regions around the world, leveraging the technological capabilities and global network cultivated over a history of more than 50 years, we discover, propose, and implement cutting-edge technologies such as AI, IoT, and autonomous driving. <https://www.macnica.co.jp/en/>

About Fixstars Corporation

Fixstars is a technology company dedicated to accelerating AI inference and training through advanced software optimization solutions. It supports innovation in healthcare, manufacturing, finance, mobility, and other industries. For more information, visit: <https://www.fixstars.com/>

Media Contact

Public Relations, Fixstars Corporation

Email: press@fixstars.com

Tel: +81-3-6420-0751

*Company names, product names, and service names mentioned in this document are trademarks or registered trademarks of their respective companies.