Intel launches Gaudi 3 AI chip, 50% faster than NVIDIA's H100

Intel has just officially announced the latest AI processing chip line called Gaudi 3, with the goal of challenging the leading position that NVIDIA currently holds in the AI ??industry.

Intel has just officially announced the latest AI processing chip line called **Gaudi 3** , with the goal of challenging the leading position that NVIDIA currently holds in the AI ??industry. The new chip promises to significantly improve performance and efficiency compared to NVIDIA's H100 GPU, and even AMD's MI300X.

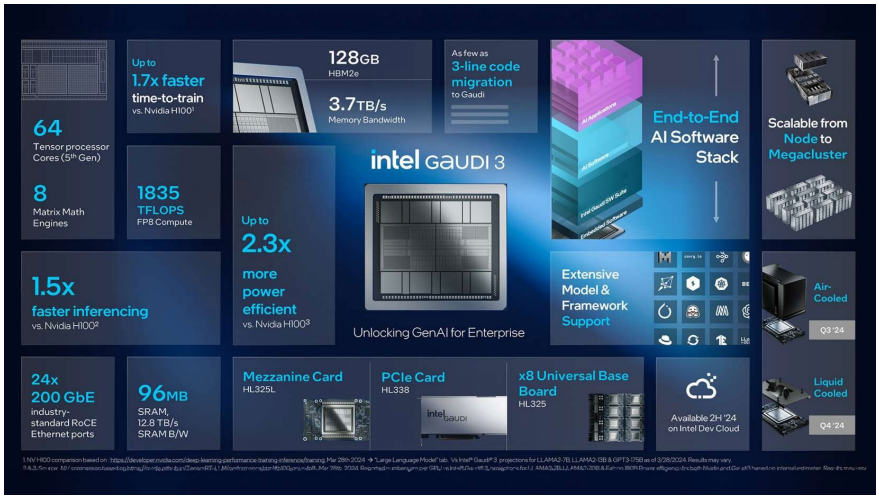
According to Intel's statement, Gaudi 3 saves twice as much energy and can perform AI tasks one and a half times faster than H100. These impressive numbers could theoretically help significantly reduce the already extremely expensive costs associated with training large AI models. The chip also comes with various modules, including an 8-chip configuration on the motherboard and a card compatible with existing Intel server designs.

Intel said it tested the chip on popular AI models from Meta's Llama 2 and Falcon projects. The results show that Gaudi 3 has good training ability for AI models, while also meeting the ability to effectively run complex applications.

According to Intel's internal benchmark data, Gaudi 3 is on average 50% faster than H100 when training large language models such as Llama with 7 billion and 13 billion parameters, as well as GPT-3 with 175 billion parameters . Additionally, Gaudi 3 is expected to achieve 40% better power efficiency than H100 in inference tasks.

'Compared to Nvidia H100, **Intel Gaudi 3** is expected to deliver 50% faster average training times on the Llama2 models with 7B and 13B specs as well as the GPT-3 175B spec model. Additionally, the inference throughput of the Intel Gaudi 3 accelerator is predicted to outperform the H100 by an average of 50%1 and 40% for average inference power efficiency on the Llama 7B and 70B parameters, also as the Falcon 180B' parametric models.

Gaudi 3's full specifications details are listed in the image below:



Gaudi 3 connects to the host CPU via a high-speed PCIe Gen 5 x16 link. Intel is also developing liquid-cooled variants for even higher performance.

The launch of Gaudi 3 is Intel's first serious challenge to NVIDIA's dominance, which is estimated to account for about 80% of the global AI chip market. With customers like Dell, HP and Supermicro building servers with the new chips, Intel can be confident of gaining a significant market share in the near future. Gaudi 3 will be available to major Intel OEMs – including Dell Technologies, HPE, Lenovo and Supermicro – in Q2 2024.