For some Chinese chips, "no end in sight" to support the full-parameter version of DeepSeek

*Note: These are Jeffrey Ding's informal and unofficial translations -- all credit for the original goes to the authors and the original text linked below. These are informal translations and all credit for the original work goes to the authors. Others are welcome to share **excerpts** from these translations as long as my original translation is cited. Commenters should be aware that the Google Doc is also publicly shareable by link. These translations are part of the ChinAI newsletter - weekly-updated library of translations from Chinese thinkers on AI-related issues: https://chinai.substack.com/*

Author: Yonggang Bao
Source: AItechtalk
Date: March 2, 2025
Original Mandarin: https://mp.weixin.qq.com/s/BiieiP-ViGXzH_5j9xLOyA

Unlike the lively scene of nearly 20 AI chip companies busy announcing the completion of their support for the DeepSeek distilled model versions just after the Spring Festival holiday, only a few companies announced the completion of their support for the full-HP version of DeepSeek model half a month later, which also truly reflects the true strength of domestic AI chips.

"As long as the manufacturer has supported the training and inference of large models before, there will be no difficulty in adapting DeepSeek." AI chip software engineer Zihao said, "Our company's application engineers (AE) can straightforwardly complete the adaptation of the DeepSeek distillation model."

This is enough to explain why some chip companies can complete the adaptation of DeepSeek's distilled versions in a few hours, but for AI chip companies that have been committed to making large chips, adapting to the full HP version of DeepSeek can better reflect its value.

At present, Huawei, Cambricon, Moore Threads and Kunlun Chip have all publicly stated that they have completed the adaptation of the full-HP version of DeepSeek model.

"Even the chip companies that have announced the adaptation of the full-HP version of DeepSeek, their performance is not very good." Jack, a senior AI chip engineer, said, "From a technical point of view, it is only a matter of time for companies that have already (developed chips to) run large models, such as Enflame, Shanghai Biren, and Iluvatar CoreX [天数智芯], to adapt to the full-blood version of DeepSeek. Companies that have not deployed large models before may have no end in sight when it comes to supporting the full-blood version of DeepSeek."

So will the adaptation of DeepSeek's distilled versions and full-blood version become a watershed moment for AI chip companies? Why do some people say that people in domestic AI chip companies don't understand AI? Can't DeepSeek's explosive popularity at least support the listing of a domestic AI chip company?

1. The distilled DeepSeek models are just an appetizer

Last month, amidst the overwhelming news about chip companies adapting to DeepSeek, some companies clearly stated that they were supporting the distilled models, while others only said that they were adapting to DeepSeek. However, there is a huge difference between supporting the distilled model and the full-blooded model.

The full-blooded model refers to the full-parameter model of V3 and R1 that is consistent with the performance described by DeepSeek's official website. Its parameters are as high as 671B, and generally multiple high-performance GPU servers are required in parallel to run the inference service smoothly.

The distilled DeepSeek model uses the data generated by DeepSeek-R1 to fine-tune other models. The parameters range from a few billion to dozens of billions, such as DeepSeek-R1-Distill-Qwen-1.5B/7B/14B/32B, DeepSeek R1-Distill-Llama-8B/70B. The effects of these distilled models are worse than the full-blooded version, but they are easier to deploy.

"I once thought that adapting a distilled version of DeepSeek model was not very valuable, and many engineers also preferred the full-blooded version of DeepSeek, but now my thoughts have changed." Bo Lin, who has more than 20 years of experience in the chip industry, said, "The distilled version of the model can meet the chat needs of ordinary users, which is of great significance to the dissemination of AI."

Jack also said that although the accuracy of the distilled model is not as good as the full-blooded model, the distilled model can take the capabilities of the end-side AI to a higher level. The end-side resources are limited. With a distilled version of DeepSeek, for example, a particular application scenario that could only deploy a 7B model before can now achieve the performance of a 14B model.

It is not difficult for domestic AI chips to support distilled versions of DeepSeek.

Whether using the GPGPU architecture or the dedicated ASIC architecture, AI chip companies have quickly integrated support for DeepSeek. "After GPT became popular, all companies tried to support large models. DeepSeek is not fundamentally different from the previous large models. With the previous work of adapting large models, adapting the distilled versions of DeepSeek is not a problem." Zihao said.

"The CUDA-compatible GPGPU framework is indeed easier (to run DeepSeek models), but as long as ASICs are given more time to achieve their ultimate performance, their performance can exceed that of GPUs," Jack believes.

In the long run, no matter what architecture the chip is, if it only supports a few limited models, you can eventually figure out the optimal solution. DeepSeek is popular, and the mainstream models are DeepSeek and Llama and a few others. From this perspective, it is a good thing for AI chip companies.

For intelligent computing centers that use domestic AI chips, the popularity of DeepSeek is also a major benefit.

"After DeepSeek exploded, we wanted to adapt it with a card from a domestic AI chip company." Boyuan, a practitioner at a Chinese intelligent computing center, said, "But the reality is that if the (inference) performance of DeepSeek on an A100 is 100 points, this domestic card only provides a few points of performance, and even if it is optimized, it only has a performance of around 10 percent that of the A100."

Since the distilled version of DeepSeek has great value from the perspective of popularizing AI and adaptation, why do we need to adapt the full-blood version of DeepSeek?

"Only after deploying the full-blood version of DeepSeek model can one obtain the distilled version. I think this is an important reason for deploying the full-blood version of DeepSeek model." Jack said.

2. The earliest that leading Chinese AI chips can adapt to the "good" full-HP model is the end of the month

However, if you want to deploy the full-blood version of DeepSeek-R1 model with parameters as high as 671B, even the size of the Int8 precision model is as high as 671G. Calculated based on 96G HBM per card, the total size of 8 cards on a single machine is 768GB, which is barely enough to deploy the full-blood version of DeepSeek. As long as the model accuracy is higher than Int8, a single server cannot deploy the full-blood version of DeepSeek model.

At this time, multi-machine interconnection is required, which is exactly the problem that domestic AI chip companies have not yet solved well.

"Nvidia has NVLink, and domestic chips do not have a solution for connections across multiple servers. They will choose InfiniBand (IB) or high-speed Ethernet RoCE to achieve interconnection. The communication delay of these solutions is very large, which will greatly affect the final deployment effect." Jack said, "Multi-card and multi-machine interconnection is the first difficulty for domestic chips to adapt to the full-blooded version of DeepSeek. If the communication problem is not solved before, it will be difficult to do it, and there may be no end in sight to successfully adapting to the full-parameter version of DeepSeek."

Zihao believes that Moore Threads and MetaX have certain advantages in multi-machine interconnection.

Another difficulty is DeepSeek's MoE hybrid expert system. MoE is the calculation of an additional router (routing module). It will route the token to the appropriate expert weight for calculation. This routing is dynamic, which is different from the previous deployment of the Transformer large model. This is also a new challenge.

For all domestic AI chips, another flaw is that they do not natively support the FP8 data representations. The DeepSeek model uses FP8 mixed precision training. NVIDIA, the world's leading AI chip company, has natively supported FP8 since H100, and AMD MI325X has natively supported FP8.

"Not natively supporting FP8 does not mean that the full version of DeepSeek cannot be deployed, but it will bring efficiency problems. For example, using FP16 to deploy requires twice the storage." Jack said that this means that more cards are needed, and the problem is interconnections within and between server racks.

It should be noted that even the new generations of Chinese AI cards launched in 2024 do not support FP8.

According to Bo Lin, the fact that the latest domestic AI chips do not support non-IEEE defined data types such as FP8 and FP4 shows that there is no cutting-edge research within the company to guide company strategy. And NVIDIA's H100, launched in 2022, already supports FP8, and some people have already made products, so it is not difficult to "copy" them. This shows that many people who make AI chips in China do not understand AI.

Even if the technical difficulties are solved and the full-blooded version of DeepSeek can be deployed, there is still a long way to go from being “able to use” toward “convenient to use”. Jack has a deep feeling about this. When adapting to large models before, Jack’s company did resolve the inter-server interconnection problems, but it was very difficult to achieve performance improvement.

This is also the headache for domestic chip companies to adapt to the full-blooded version.

Leiphone learned that the current leading Chinese AI chip companies have only achieved 10 tokens/s with 4 servers (32 cards, FP16 data type) or 2 servers (16 cards, Int8 data type) to adapt to the full-parameter version of DeepSeek. Their goals are to reach 25 tokens/s by the end of February, which is about 25% of NVIDIA H100 (in terms of performance level)

There are also reports that Chinese listed AI chip companies have achieved 25 tokens/s performance in intelligent computing centers for deploying the full-blooded version of DeepSeek.

From the user's perspective, there are two very intuitive indicators for a better user experience using the full-blooded version of DeepSeek, one is the first word latency, and the other is the throughput per second. Generally speaking, the first word delay of 1-1.4 seconds is acceptable to most users, and generating 20 tokens per second can meet the needs of normal reading.
In this way, even the leading domestic companies will have to wait until the end of ~~February~~ March to achieve a relatively satisfactory user experience.

As for other AI chip companies, Leiphone learned that several AI chip companies in the listing guidance process have adapted to the full-blooded version of DeepSeek at a speed of 10 tokens/s or less.

Zhang Wei, who is based at a large AI chip company, judged that half of the AI chip companies may not be able to adapt to the full-blooded version of DeepSeek in the next month. Bo Lin believes that domestic AI chips will gradually support the full-blooded version of DeepSeek in the next quarter.

"It is only a matter of time before other chip companies that have successfully deployed large models adapt to the full-blooded version of DeepSeek." Jack said, "Several of these companies are in the stage of listing guidance. I think whoever can support the full-blooded version of DeepSeek faster and better will greatly increase their chances of listing, because many institutions and companies are actively deploying the full-blooded version of DeepSeek, which is conducive to AI chip companies to achieve real results and support their listing."

However, two chip investors told Leiphone that the factors for the successful listing of A-shares are relatively complex. Being able to support the full-blooded version of DeepSeek is indeed a reflection of strength, but it is difficult to say that it has a direct benefit for the final successful listing.

There is no doubt that DeepSeek is a huge benefit to domestic chips, intelligent computing centers, and AI applications. We are already on the eve of AI transformation.

Note: Zihao, Jack, Bo Lin, Boyuan, and Zhang Wei in the article are all pseudonyms.