

DeepSeek tuyên bố chip AI của Huawei đạt hiệu suất suy luận tới 60% so với NVIDIA: Cơ hội để Trung Quốc giảm phụ thuộc vào GPU từ Mỹ?

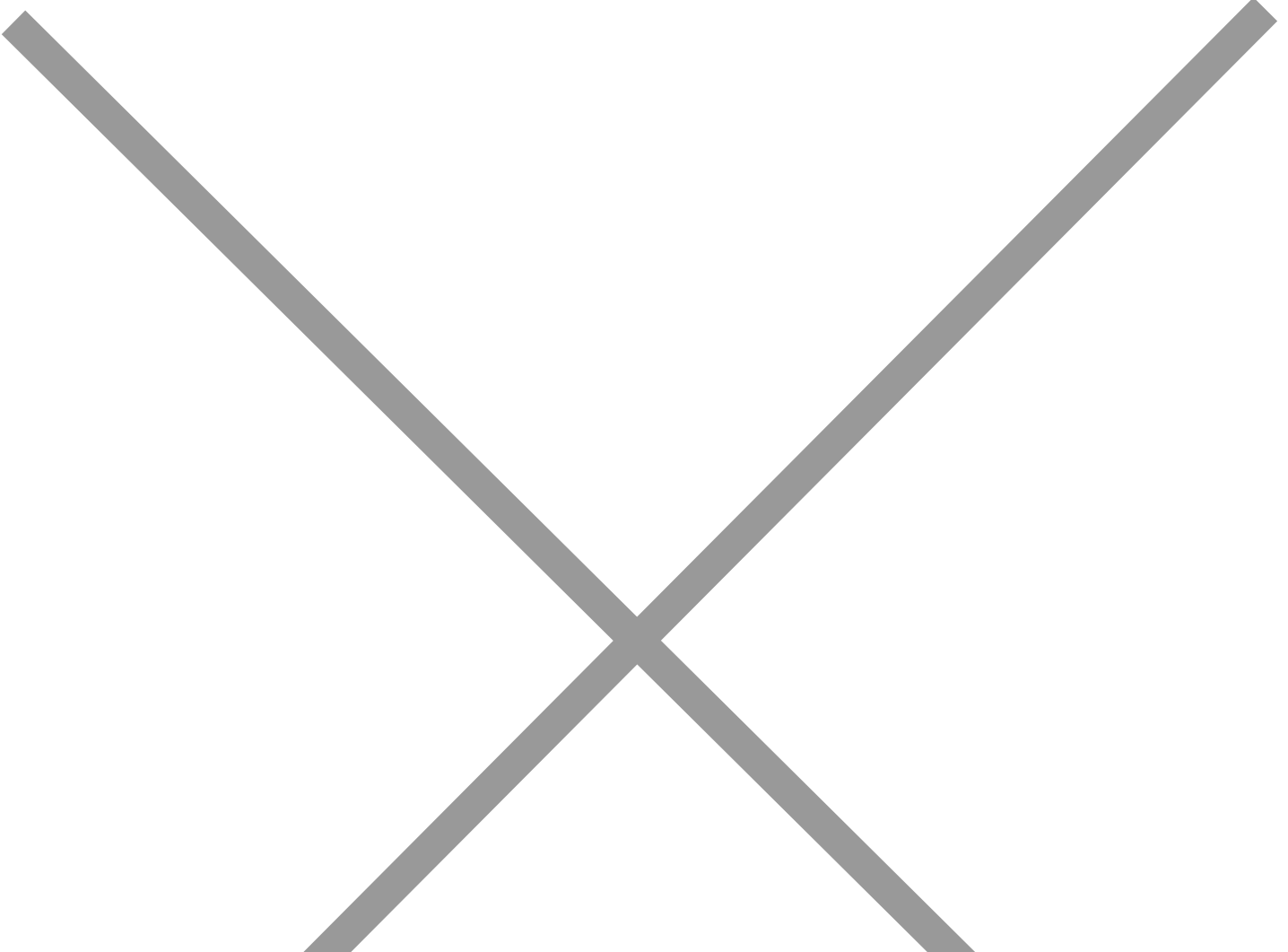
14:39 05/02/2025

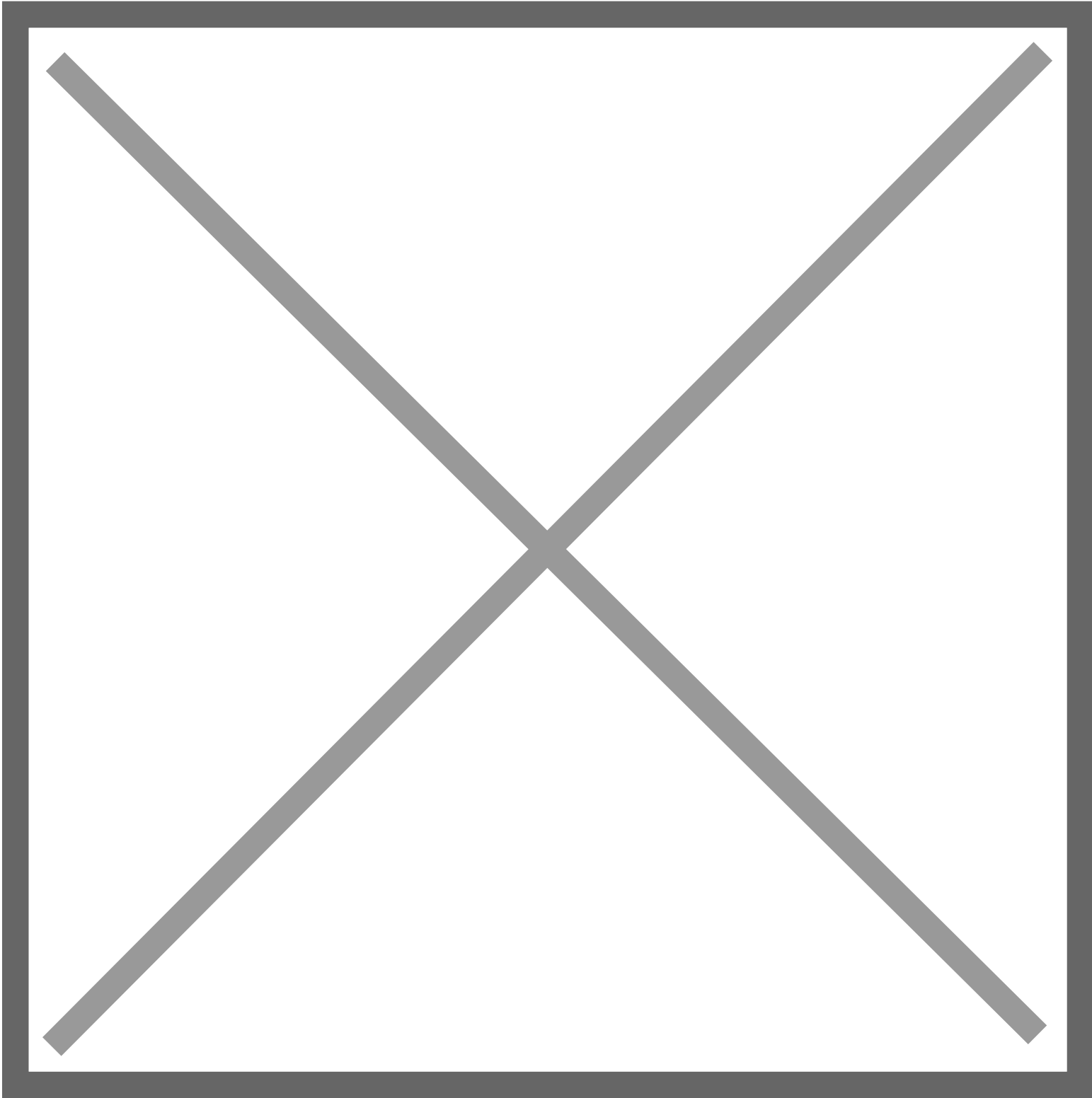
Tác giả: Đang cập nhật

Dù chưa thể cạnh tranh trong mảng đào tạo AI, con chip này cho thấy tiềm năng giúp Trung Quốc giảm phụ thuộc vào GPU NVIDIA, đặc biệt khi kết hợp với những cải tiến phần mềm từ DeepSeek.

Bộ xử lý HiSilicon Ascend 910C của Huawei là phiên bản cải tiến của Ascend 910, con chip AI ra mắt vào năm 2019. Trong bối cảnh các mô hình AI ngày càng lớn và phức tạp, hiệu suất của Ascend 910 hiện không còn đủ để đào tạo các mô hình AI tiên tiến một cách tối ưu về chi phí. Tuy nhiên, khi xét về khả năng suy luận (inference), bộ xử lý này đạt 60% hiệu suất của Nvidia H100, theo nghiên cứu của DeepSeek. Dù không phải là con chip AI mạnh nhất, Ascend 910C có thể đóng vai trò quan trọng trong việc giúp Trung Quốc giảm sự phụ thuộc vào GPU Nvidia.

Image not found or type unknown





Huawei cải thiện hiệu suất suy luận với Ascend 910C

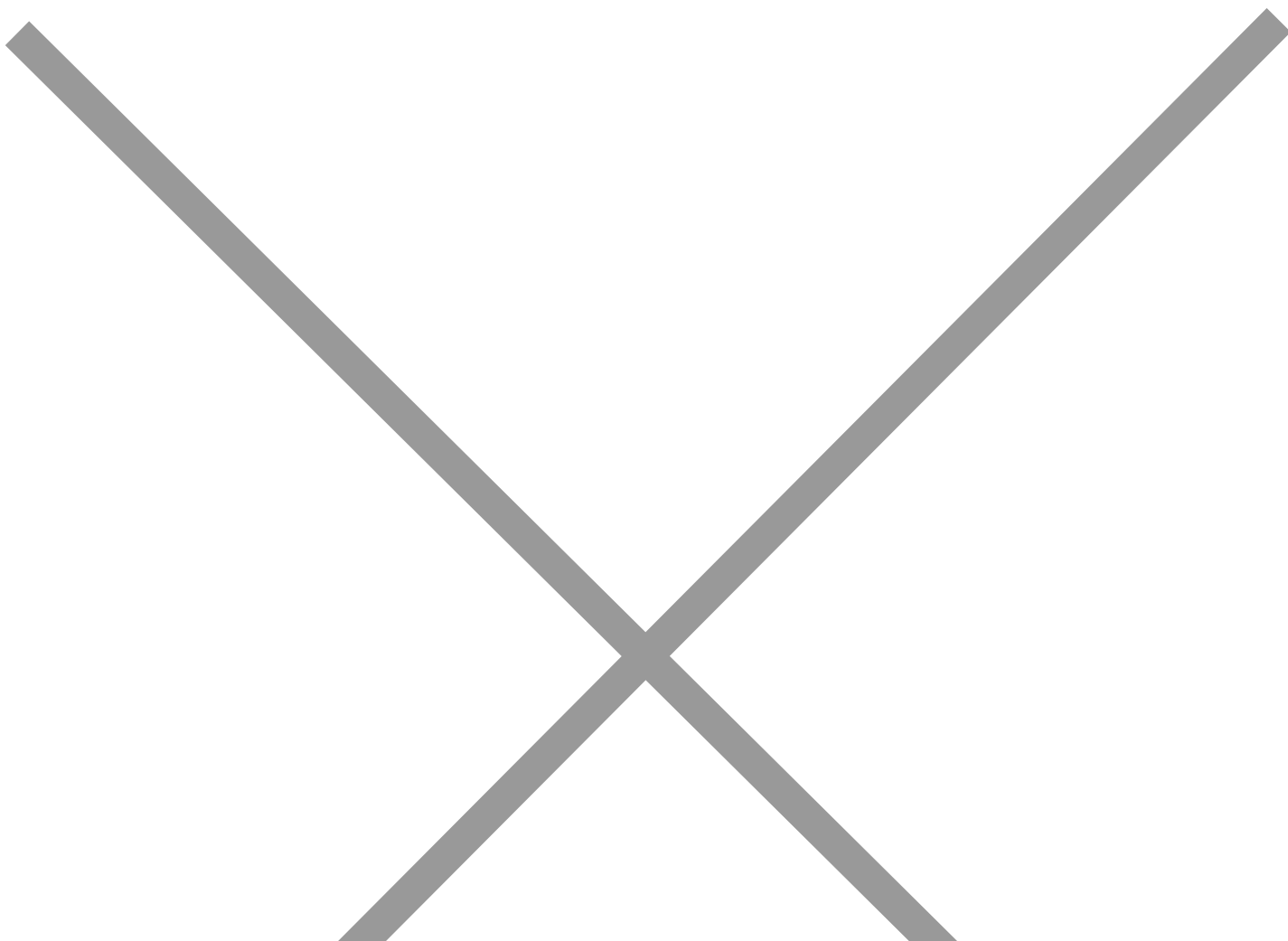
Các thử nghiệm của DeepSeek cho thấy Ascend 910C vượt mong đợi về khả năng suy luận, và nếu được tối ưu thủ công với các kernel CUNN, hiệu suất có thể còn được cải thiện hơn nữa. Nhờ sự hỗ trợ nguyên bản của DeepSeek dành cho bộ xử lý Ascend cùng kho lưu trữ PyTorch, việc chuyển đổi từ CUDA sang CUNN trở nên đơn giản, giúp tích hợp phần cứng của Huawei vào quy trình AI một cách dễ dàng.

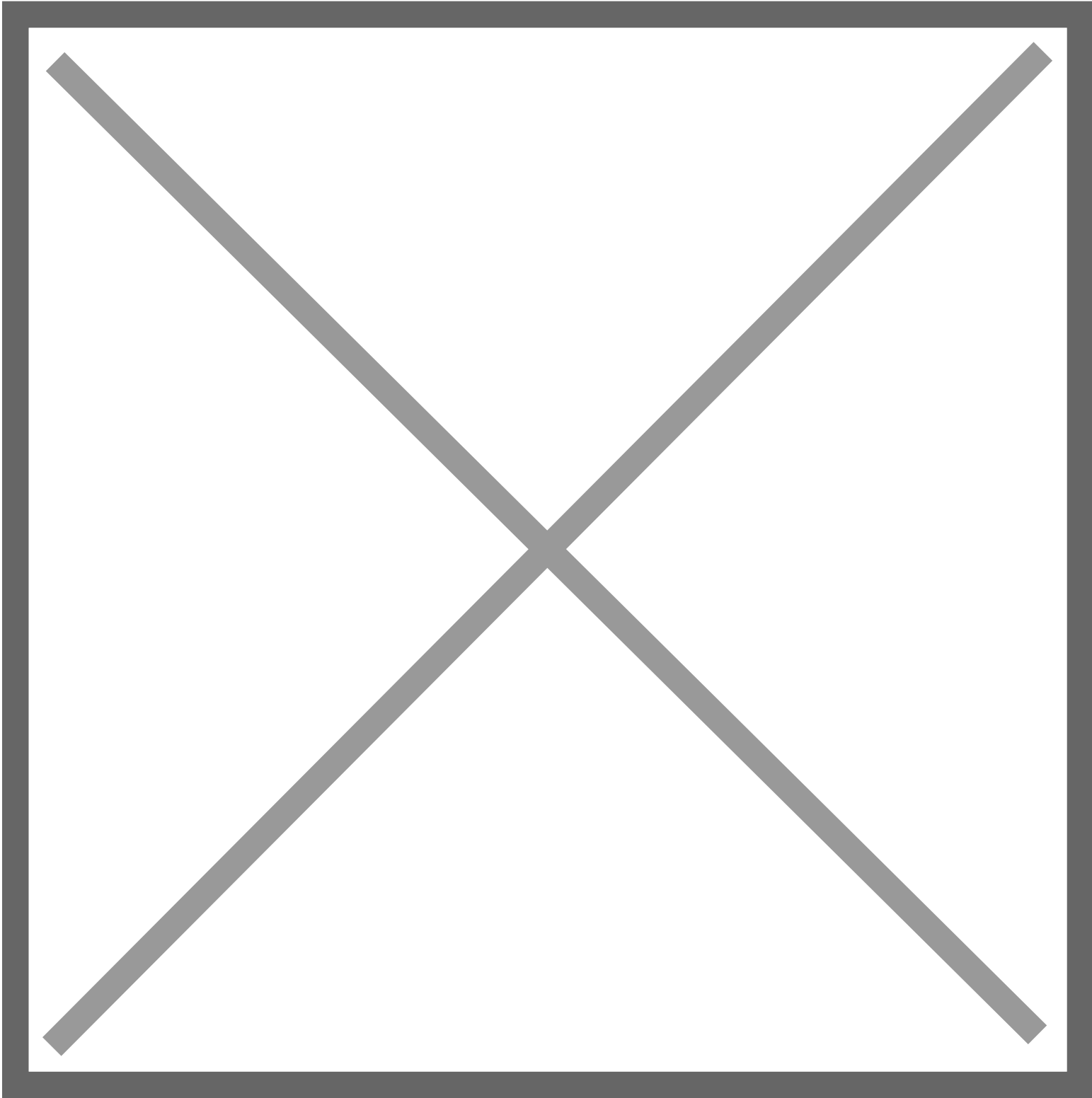
Bất chấp các lệnh trừng phạt từ Mỹ và hạn chế trong tiếp cận công nghệ sản xuất chip tiên tiến của TSMC, Huawei vẫn đang tiến bộ nhanh chóng trong lĩnh vực chip AI. SMIC, đối tác sản xuất của Huawei, đã bắt kịp trình độ của TSMC giai đoạn 2019–2020 và cho ra đời một con chip có thể cạnh tranh với Nvidia A100 và H100.

Hạn chế của Ascend 910C: Đào tạo AI vẫn là thế mạnh của Nvidia

Dù đạt kết quả khả quan trong suy luận, Ascend 910C chưa phải là lựa chọn lý tưởng cho đào tạo AI, lĩnh vực mà Nvidia vẫn giữ vững vị thế dẫn đầu. Theo ông Yuchen Jin, đại diện của DeepSeek, tính ổn định khi đào tạo dài hạn là điểm yếu lớn nhất của các bộ xử lý AI Trung Quốc. Nguyên nhân đến từ việc Nvidia đã phát triển hệ sinh thái phần cứng và phần mềm trong suốt hơn hai thập kỷ, tạo ra lợi thế vượt trội trong lĩnh vực này.

image not found or type unknown





Ascend 910C tiếp tục sử dụng thiết kế chiplet, với SoC tính toán chính chứa khoảng 53 tỷ bóng bán dẫn. Trong khi Ascend 910 đời đầu sử dụng chiplet tính toán do TSMC sản xuất trên tiến trình N7+ (7nm với EUV), Ascend 910C chuyển sang tiến trình N+2 (thế hệ thứ hai của 7nm) của SMIC.

Tương lai của AI: Liệu Huawei có thể thoát khỏi sự thống trị của Nvidia?

Một số chuyên gia dự đoán rằng khi các mô hình AI tiếp tục hội tụ vào kiến trúc Transformer, hệ

sinh thái phần mềm của Nvidia có thể mất đi phần nào lợi thế. Đồng thời, việc DeepSeek tập trung vào tối ưu hóa phần cứng và phần mềm có thể giúp giảm sự phụ thuộc vào GPU của Nvidia, mang lại một lựa chọn tiết kiệm chi phí hơn, đặc biệt trong lĩnh vực suy luận AI.

Tuy nhiên, để có thể cạnh tranh trên quy mô toàn cầu, Trung Quốc cần giải quyết bài toán về độ ổn định khi đào tạo mô hình AI và tiếp tục nâng cấp cơ sở hạ tầng tính toán AI. Nếu vượt qua rào cản này, Huawei có thể trở thành một đối thủ đáng gờm trong lĩnh vực phần cứng AI, mở ra cơ hội thay thế GPU Nvidia trong nhiều ứng dụng chiến lược.

Link bài viết: <http://www.tinthat.vn/deepseek-tuyen-bo-chip-ai-cua-huawei-dat-hieu-suat-suy-luan-toi-60-so-voi-nvidia-co-hoi-de-trung-quoc-giam-phu-thuoc-vao-gpu-tu-my>