XU HƯỚNG VÀ ẢNH HƯỞNG CỦA KIẾN TRÚC PHẦN CỨNG ĐỐI VỚI SỰ PHÁT TRIỂN CÁC MÔ HÌNH TRÍ TUỆ NHÂN TẠO (AI) QUY MÔ LỚN

Nguyễn Văn Thuận
Trường Đại học Tiền Giang
Email:nguyenvanthuan@tgu.edu.vn.

Tóm tắt: *Bài viết này nghiên cứu về ảnh hưởng và xu hướng phần cứng đối với sự phát triển của các mô hình trí tuệ nhân tạo (AI) quy mô lớn trong những năm gần đây. Quá trình huấn luyện và triển khai các mô hình này đòi hỏi năng lực tính toán cực lớn, vượt xa khả năng của các kiến trúc phần cứng truyền thống. Do đó, kiến trúc phần cứng đã trở thành một yếu tố then chốt, ảnh hưởng trực tiếp đến hiệu năng, khả năng mở rộng, chi phí và hiệu quả năng lượng của các hệ thống AI quy mô lớn. Bài viết cũng tập trung phân tích vai trò và ảnh hưởng của các kiến trúc phần cứng hiện đại đối với sự phát triển của các mô hình trí tuệ nhân tạo quy mô lớn, đồng thời tổng hợp và đánh giá các xu hướng phần cứng nổi bật hiện nay, bao gồm bộ tăng tốc AI chuyên biệt, kiến trúc lấy bộ nhớ làm trung tâm và hệ thống tính toán phân tán hiệu năng cao. Trên cơ sở đó, tác giả thảo luận về triển vọng, các thách thức còn tồn tại và đề xuất một số định hướng nghiên cứu nhằm thúc đẩy sự phát triển bền vững của trí tuệ nhân tạo trong tương lai.*

Từ khóa: *Kiến trúc phần cứng, trí tuệ nhân tạo quy mô lớn, bộ tăng tốc AI, GPU, mô hình ngôn ngữ lớn.*
Nhận bài: 12/01/2026; Biên tập: 13/01/2026; Phản biện: 19/01/2026; Duyệt đăng: 26/01/2026.

1. Đặt vấn đề

Trong thập kỷ qua, trí tuệ nhân tạo (AI) đã phát triển mạnh mẽ với sự ra đời của các mô hình quy mô lớn có hàng nghìn tỷ tham số. Những mô hình này thể hiện năng lực vượt trội trong học sâu, suy luận và tổng quát hóa, mở ra nhiều ứng dụng thực tiễn. Thành công của AI quy mô lớn không chỉ phụ thuộc vào dữ liệu và thuật toán mà còn gắn chặt với năng lực tính toán của hệ thống phần cứng.

Khi kích thước mô hình, dữ liệu huấn luyện và tài nguyên tính toán cùng mở rộng, nhu cầu về kiến trúc phần cứng hiệu năng cao trở nên cấp thiết. Các hệ thống cần khả năng xử lý song song mạnh mẽ, băng thông bộ nhớ lớn và hiệu quả năng lượng tối ưu. Trong bối cảnh đó, kiến trúc dựa trên CPU truyền thống dần bộc lộ hạn chế trước yêu cầu tính toán phức tạp của AI hiện đại.

Để đáp ứng nhu cầu này, nhiều giải pháp phần cứng mới đã được phát triển như GPU, FPGA, TPU, ASIC và các hệ thống tính toán phân tán. Phần cứng không chỉ là nền tảng thực thi mà còn ảnh hưởng trực tiếp đến thiết kế mô hình, thuật toán huấn luyện và chiến lược triển khai, hình thành xu hướng đồng thiết kế giữa phần cứng và thuật toán. Tuy nhiên, vẫn thiếu các nghiên cứu tổng hợp hệ thống về tác động và xu hướng phát triển kiến trúc phần cứng đối với AI quy mô lớn. Vì vậy, bài báo phân tích ảnh hưởng của phần cứng đến sự phát triển AI, đồng thời tổng hợp các xu hướng nổi bật và thảo luận định hướng nghiên cứu trong tương lai.

2. Nội dung nghiên cứu

2.1. Các mô hình trí tuệ nhân tạo quy mô lớn

2.1.1. Khái niệm và đặc điểm cơ bản

Các mô hình trí tuệ nhân tạo quy mô lớn (Large-Scale Artificial Intelligence Models) là lớp mô hình học máy được đặc trưng bởi số lượng tham số rất lớn, thường dao động từ hàng trăm triệu đến hàng nghìn tỷ tham số. Những mô hình này được huấn luyện trên các tập dữ liệu khổng lồ và có khả năng học biểu diễn phức tạp, cho phép thực hiện nhiều nhiệm vụ khác nhau với hiệu năng vượt trội so với các mô hình truyền thống. Tiêu biểu cho nhóm này là các mô hình dựa trên kiến trúc Học chuyển đổi(Transformer), trong lĩnh vực xử lý ngôn ngữ tự nhiên, thị giác máy tính và mô hình đa phương thức.

Một điểm quan trọng của các mô hình AI quy mô lớn là tính tổng quát cao và khả năng chuyển giao tri thức. Thay vì được huấn luyện cho một nhiệm vụ cụ thể, các mô hình này thường được huấn luyện trước (pre-training) trên dữ liệu tổng quát, sau đó tinh chỉnh (fine-tuning) cho các ứng dụng khác nhau. Điều này làm thay đổi cách tiếp cận truyền thống trong phát triển hệ thống AI, Ngoài ra, các mô hình AI quy mô lớn thường có cấu trúc sâu, độ phức tạp cao và yêu cầu khả năng tính toán song song mạnh mẽ. Sự gia tăng quy mô mô hình không chỉ mang lại lợi ích về độ chính xác mà còn kéo theo nhiều thách thức về mặt kỹ thuật, đặc biệt là liên quan đến hạ tầng phần cứng và hệ thống tính toán.

<table>
<tr><th colspan="8">Nền tảng của các mô hình Trí tuệ Nhân tạo quy mô lớn</th></tr>
<tr><th colspan="2">Bộ dữ liệu</th><th colspan="4">Kiến trúc thuật toán</th><th>Tài nguyên tính toán</th><th>Khung học sâu</th></tr>
<tr><td rowspan="2">Dữ liệu công khai</td><td rowspan="3">Dịch vụ xử lý bộ dữ liệu công cộng ces</td><td colspan="2">Học sâu</td><td colspan="2">Học chuyển đổi</td><td>Nền tảng tính toán</td><td>TensoftFlow (thư viện mã nguồn mở dung cho máy học)</td></tr>
<tr><td rowspan="2">Dữ liệu riêng tư</td><td>Học đối kháng</td><td>Học tăng cường</td><td>Học không giám sát</td><td>Học ít dữ liệu</td><td rowspan="2">Lập lịch tính toán</td><td rowspan="2">PyTorch (thư viện học máy)</td></tr>
</table>

Hình 1. Nền tảng của các mô hình trí tuệ nhân tạo quy mô lớn

2.1.2. Yêu cầu về tài nguyên tính toán và bộ nhớ

Huấn luyện các mô hình trí tuệ nhân tạo quy mô lớn đòi hỏi nguồn tài nguyên tính toán rất lớn, với hàng tỷ phép toán ma trận và vectơ, yêu cầu khả năng xử lý song song cao và băng thông bộ nhớ lớn. Do đó, CPU truyền thống thường không đáp ứng được yêu cầu hiệu năng. Bên cạnh năng lực tính toán, bộ nhớ giữ vai trò then chốt, không chỉ để lưu trữ tham số mô hình mà còn chứa dữ liệu trung gian trong quá trình lan truyền xuôi và truyền ngược. Khi mô hình mở rộng, nhu cầu bộ nhớ tăng mạnh, dễ gây nghẽn cổ chai về băng thông. Vì vậy, các công nghệ như bộ nhớ băng thông cao và kiến trúc lấy bộ nhớ làm trung tâm được phát triển. Ngoài ra, huấn luyện AI quy mô lớn thường dựa trên hệ thống phân tán với nhiều bộ tăng tốc hoạt động song song, đòi hỏi kết nối tốc độ cao và cơ chế đồng bộ hiệu quả giữa các nút tính toán.

2.2. Kiến trúc phần cứng phục vụ trí tuệ nhân tạo

2.2.1. Kiến trúc CPU và vai trò truyền thống trong trí tuệ nhân tạo

Bộ xử lý trung tâm (CPU) là nền tảng tính toán truyền thống trong các hệ thống máy tính và đóng vai trò quan trọng trong giai đoạn đầu phát triển của trí tuệ nhân tạo. CPU được thiết kế để xử lý linh hoạt nhiều loại tác vụ khác nhau, với khả năng điều khiển luồng và xử lý logic phức tạp. Trong các hệ thống AI, CPU thường đảm nhiệm các nhiệm vụ như tiền xử lý dữ liệu, điều phối luồng tính toán và quản lý hệ thống. Tuy nhiên, do số lượng lõi xử lý hạn chế và khả năng xử lý song song thấp so với các kiến trúc chuyên biệt, CPU không phù hợp cho các tác vụ tính toán cường độ cao như huấn luyện mô hình học sâu quy mô lớn. Khi kích thước mô hình và dữ liệu tăng lên, CPU dần trở thành nút thắt cổ chai về hiệu năng, buộc các hệ thống AI hiện đại phải chuyển sang các kiến trúc phần cứng khác có khả năng tính toán song song cao hơn.

Loại kiến trúc phần cứng	Đặc điểm chính	Vai trò trong AI quy mô lớn	Ưu điểm	Hạn chế	Ứng dụng tiêu biểu
CPU đa lõi	Kiến trúc xử lý tuần tự, linh hoạt, hỗ trợ nhiều tác vụ	Điều phối hệ thống, tiền xử lý dữ liệu, điều khiển huấn luyện	Linh hoạt, dễ lập trình	Hiệu năng thấp cho tính toán song lớn	Pipeline huấn luyện, inference nhẹ
GPU (Graphics Processing Unit)	Hàng nghìn lõi song băng thông bộ nhớ cao	Nền tảng chính cho huấn luyện và suy luận AI quy mô lớn	Hiệu năng cao, hệ sinh thái phần mềm mạnh	Tiêu thụ năng lượng lớn	Huấn luyện LLM, thị giác máy tính
TPU (Tensor Processing Unit)	Gia tốc chuyên dụng cho tensor và phép nhân m trận	Tối ưu cho học sâu và Transformer	Hiệu quả năng lượng cao	Pụ thuộc nền tảng, ít linh hoạt	Huấn luyện mô hình cực lớn
FPGA	Có thể tái cấu hình phần cứng	Tối ưu inference và ứng dụng chuyên biệt	Tiết kiệm năng lượng, độ trễ thấp	Lập trình phức tạp	Edge AI, inference thời gian thực
ASIC AI Accelerator	Thiết kế chuyên dụng cho AI	Tối ưu hiệu suất/năng lượng	Hiệu suất cao nhất	Không linh hoạt, chi phí cao	Trung tâm dữ liệu AI
Bộ nhớ tốc độ cao (HBM, DDR5)	Băng thông lớn, độ trễ thấp	Giảm nghẽn cổ chai dữ liệu	Cải thiện hiệu năng mô hình lớn	Giá thành cao	LLM, mô hình đa phương thức
Hệ thống liên kết (NVLink, InfiniBand)	Kết nối tốc độ cao giữa các thiết bị	Huấn luyện phân tán quy mô lớn	Khả năng mở rộng tốt	Chi phí hạ tầng	Data center AI
Hệ thống lưu trữ phân tán	Lưu trữ dữ liệu và checkpoint lớn	Hỗ trợ huấn luyện dài hạn	Dung lượng lớn	Độ trễ I/O	Huấn luyện LLM

Hình 2. Mô tả các kiến trúc phần cứng phục vụ trí tuệ nhân tạo quy mô lớn

2.2.2. Kiến trúc GPU và tính toán song song cho AI

Bộ xử lý đồ họa (GPU) đã trở thành kiến trúc phần cứng chủ đạo trong lĩnh vực trí tuệ nhân tạo, học sâu. GPU được thiết kế với hàng nghìn lõi xử lý, cho phép thực hiện song song số lượng lớn phép toán số học, các phép nhân ma trận và vectơ, vốn là nền tảng của các mô hình học sâu.

Ưu điểm nổi bật của GPU nằm ở khả năng tận dụng tính song song dữ liệu ở quy mô lớn, giúp rút ngắn đáng kể thời gian huấn luyện mô hình AI. Bên cạnh đó, sự phát triển của các nền tảng phần mềm hỗ trợ như CUDA (Compute Unified Device Architecture), các thư viện học sâu đã góp phần thúc đẩy việc ứng dụng GPU trong nghiên cứu và triển khai AI.

2.2.3. FPGA và khả năng tái cấu hình cho trí tuệ nhân tạo

FPGA (Field-Programmable Gate Array) là một dạng kiến trúc phần cứng có khả năng tái cấu hình linh hoạt, cho phép người dùng tùy chỉnh mạch logic theo yêu cầu ứng dụng. Trong lĩnh vực trí tuệ nhân tạo, FPGA được sử dụng chủ yếu cho các tác vụ suy luận (inference), nơi yêu cầu độ trễ thấp và hiệu quả năng lượng cao.

Ưu điểm của FPGA là khả năng tối ưu hóa phần cứng cho một mô hình hoặc thuật toán cụ thể, từ đó giảm tiêu thụ năng lượng và tăng hiệu suất so với các kiến trúc đa dụng. Tuy nhiên, việc lập trình FPGA phức tạp hơn so với GPU và đòi hỏi kiến thức chuyên sâu về thiết kế phần cứng, điều này hạn chế khả năng phổ biến của FPGA trong huấn luyện các mô hình AI quy mô lớn.

2.2.4. Các bộ tăng tốc AI chuyên biệt (ASIC, TPU, NPU)

Trước nhu cầu ngày càng tăng về hiệu năng và hiệu quả năng lượng, các bộ tăng tốc AI chuyên biệt đã được phát triển nhằm phục vụ trực tiếp cho các tác vụ trí tuệ nhân tạo. Các kiến trúc này, bao gồm ASIC (Application-Specific Integrated Circuit), TPU (Tensor Processing Unit) và NPU (Neural Processing Unit), được thiết kế tối ưu cho các phép toán đặc trưng của AI, đặc biệt là các phép toán tensor.

So với GPU, các bộ tăng tốc AI chuyên biệt thường đạt hiệu năng trên mỗi watt(công suất tiêu thụ điện) cao hơn và độ trễ thấp hơn, đặc biệt trong các môi trường triển khai quy mô lớn như trung tâm dữ liệu. Tuy nhiên, nhược điểm của các kiến trúc này là tính linh hoạt thấp, khó thích ứng nhanh với các thuật toán mới. Do đó, việc lựa chọn bộ tăng tốc phù hợp phụ thuộc vào mục tiêu ứng dụng và yêu cầu của hệ thống cụ thể.

2.2.5. Hệ thống tính toán phân tán và kiến trúc liên kết tốc độ cao

Đối với các mô hình trí tuệ nhân tạo quy mô lớn, một bộ xử lý đơn lẻ không còn đủ khả năng đáp ứng yêu cầu tính toán và bộ nhớ. Do đó, các hệ thống tính toán phân tán, bao gồm nhiều bộ tăng tốc AI được kết nối với nhau, đã trở thành xu hướng tất yếu. Các hệ thống này cho phép phân chia dữ liệu và mô hình trên nhiều nút tính toán, từ đó mở rộng khả năng huấn luyện và triển khai AI.

Hiệu năng của các hệ thống phân tán phụ thuộc lớn vào kiến trúc liên kết giữa các nút tính toán. Các công nghệ kết nối tốc độ cao như NVLink, InfiniBand (đường cao tốc dữ liệu) và các giao thức truyền thông chuyên biệt giúp giảm độ trễ và tăng băng thông truyền dữ liệu, đóng vai trò quan trọng trong việc duy trì hiệu năng tổng thể của hệ thống AI quy mô lớn.

2.3. Ảnh hưởng của kiến trúc phần cứng đối với các mô hình AI

2.3.1. Hiệu năng và khả năng mở rộng

Kiến trúc phần cứng đóng vai trò quyết định đối với hiệu năng và khả năng mở rộng của các mô hình trí tuệ nhân tạo quy mô lớn. Khi số lượng tham số và dữ liệu huấn luyện tăng lên nhanh chóng, khối lượng tính toán chủ yếu tập trung vào các phép toán tensor (khái niệm toán học trong lĩnh vực AI) và nhân ma trận quy mô lớn. Các kiến trúc phần cứng có khả năng tính toán song song cao, đặc biệt là GPU và các bộ tăng tốc AI chuyên biệt, đã chứng minh hiệu quả vượt trội so với CPU truyền thống trong việc rút ngắn thời gian huấn luyện. Ngoài năng lực tính toán đơn lẻ, khả năng mở rộng theo chiều ngang thông qua các hệ thống đa GPU hoặc đa bộ tăng tốc là yếu tố then chốt đối với huấn luyện AI. Tuy nhiên, hiệu năng mở rộng không chỉ phụ thuộc vào số lượng thiết bị mà còn phụ thuộc mạnh mẽ vào kiến trúc liên kết và băng thông truyền thông giữa các nút tính toán.

2.3.2. Thiết kế mô hình và thuật toán huấn luyện

Sự phát triển của kiến trúc phần cứng không chỉ ảnh hưởng đến hiệu năng thực thi mà còn tác động trực tiếp đến thiết kế mô hình và thuật toán huấn luyện. Kiến trúc Transformer, nền tảng của nhiều mô hình AI quy mô lớn hiện nay, được đánh giá là phù hợp với các kiến trúc phần cứng song song như GPU và TPU nhờ cấu trúc tính toán ma trận đồng nhất và khả năng chia nhỏ tác vụ hiệu quả. Bên cạnh đó, các chiến lược song song hóa như song song dữ liệu, song song mô hình và song song pipeline(đường ống) được thiết kế dựa trên đặc điểm của kiến trúc phần cứng mục tiêu. Việc lựa chọn và kết hợp các chiến lược này ảnh hưởng trực tiếp đến tốc độ hội tụ và khả năng mở rộng của mô hình.

2.3.3. Bộ nhớ, băng thông và hiệu quả năng lượng

Đối với các mô hình AI quy mô lớn, bộ nhớ và băng thông truyền dữ liệu thường trở thành nút thắt cổ chai nghiêm trọng. Khi kích thước mô hình vượt quá dung lượng bộ nhớ cục bộ của một thiết bị, việc truy cập bộ nhớ ngoài hoặc trao đổi dữ liệu giữa các thiết bị làm suy giảm đáng kể hiệu năng tổng thể. Do đó, các kiến trúc phần cứng hiện đại ngày càng chú trọng đến việc tích hợp bộ nhớ băng thông cao và tối ưu hóa luồng dữ liệu.

Hiệu quả năng lượng cũng là một khía cạnh quan trọng chịu ảnh hưởng trực tiếp từ kiến trúc phần cứng. Việc huấn luyện các mô hình AI tiêu thụ lượng điện năng rất lớn. Các bộ tăng tốc AI chuyên biệt như TPU và ASIC đã được chứng minh là mang lại hiệu năng cao hơn so với GPU đa dụng, đặc biệt trong các trung tâm dữ liệu.

2.4. Ảnh hưởng của kiến trúc phần cứng đối với các mô hình AI quy mô lớn

2.4.1. Chuyên biệt hóa bộ tăng tốc cho trí tuệ nhân tạo

Xu hướng nổi bật nhất trong kiến trúc phần cứng cho AI quy mô lớn là sự chuyển dịch từ các bộ xử lý đa dụng sang các bộ tăng tốc chuyên biệt.

Xu hướng chuyên biệt hóa không chỉ nhằm tăng tốc huấn luyện mà còn hướng tới tối ưu hóa giai đoạn suy luận của các mô hình AI quy mô lớn. Việc thiết kế phần cứng gắn chặt với đặc điểm của mô hình và khối lượng công việc cụ thể cho phép giảm độ trễ, tiết kiệm năng lượng và giảm chi phí vận hành trong các hệ thống được triển khai ở quy mô lớn.

2.4.2. Đồng thiết kế phần cứng - thuật toán

Sự gia tăng độ phức tạp của các mô hình AI quy mô lớn đã thúc đẩy xu hướng đồng thiết kế giữa phần cứng và thuật toán. Thay vì phát triển phần cứng và mô hình một cách độc lập, các nghiên cứu gần đây tập trung vào việc thiết kế mô hình AI sao cho phù hợp với đặc điểm của kiến trúc phần cứng mục tiêu, đồng thời điều chỉnh kiến trúc phần cứng để khai thác tối đa cấu trúc tính toán của mô hình. Xu hướng này rất rõ trong các mô hình dựa trên Transformer, nơi các phép toán ma trận lớn có thể được ánh xạ hiệu quả lên các bộ tăng tốc chuyên biệt. Đồng thiết kế giúp giảm chi phí truyền thông, tối ưu hóa sử dụng bộ nhớ và cải thiện hiệu năng tổng thể của hệ thống, đóng vai trò quan trọng trong việc mở rộng mô hình AI lên quy mô rất lớn.

2.4.3. Kiến trúc lấy bộ nhớ làm trung tâm và bộ nhớ băng thông cao

Khi quy mô mô hình AI tiếp tục tăng, bộ nhớ và băng thông truyền dữ liệu ngày càng trở thành nút thắt cổ chai chính của hệ thống. Do đó, một xu hướng quan trọng trong kiến trúc phần cứng cho AI là chuyển từ kiến trúc lấy tính toán làm trung tâm sang kiến trúc lấy bộ nhớ làm trung tâm. Các giải pháp như bộ nhớ băng thông cao (HBM), bộ nhớ xếp chồng 3D và tích hợp bộ nhớ gần đơn vị tính toán đang được áp dụng nhằm giảm độ trễ truy cập và tăng thông lượng dữ liệu.

Ngoài ra, các kỹ thuật quản lý bộ nhớ thông minh, bao gồm phân cấp bộ nhớ và tối ưu hóa luồng dữ liệu, cũng đóng vai trò quan trọng trong việc hỗ trợ huấn luyện các mô hình AI quy mô lớn. Xu hướng này cho thấy sự phát triển của AI trong tương lai sẽ phụ thuộc không chỉ năng lực tính toán mà còn khả năng xử lý và truyền tải dữ liệu hiệu quả.

2.4.4. Hệ thống tính toán phân tán quy mô lớn và kết nối tốc độ cao

Huấn luyện các mô hình AI quy mô lớn thường vượt quá khả năng của một thiết bị đơn lẻ, dẫn đến nhu cầu xây dựng các hệ thống tính toán phân tán quy mô lớn. Xu hướng hiện nay là phát triển các cụm máy tính bao gồm hàng trăm hoặc hàng nghìn bộ tăng tốc AI, được kết nối với nhau thông qua các công nghệ liên kết tốc độ cao như NVLink, InfiniBand và các giao thức truyền thông chuyên biệt. Các kiến trúc liên kết tốc độ cao giúp giảm độ trễ và chi phí truyền thông giữa các nút tính toán. Đồng thời, các framework huấn luyện phân tán ngày càng được tối ưu để tận dụng tốt hơn các kiến trúc phần cứng hiện đại.

2.4.5. Hướng tới hiệu quả năng lượng và phát triển bền vững

Một xu hướng quan trọng khác trong kiến trúc phần cứng cho AI quy mô lớn là tập trung vào hiệu quả năng lượng và phát triển bền vững. Việc huấn luyện các mô hình AI lớn tiêu thụ lượng điện năng đáng kể, đặt ra thách thức về chi phí và tác động môi trường. Do đó, các kiến trúc phần cứng mới đang được thiết kế nhằm tối ưu hóa hiệu năng, giảm tiêu thụ năng lượng và khí thải carbon. Các giải pháp bao gồm sử dụng bộ tăng tốc chuyên biệt, tối ưu hóa điện áp và tần số, cũng như kết hợp các kỹ thuật phần mềm nhằm giảm số phép toán không cần thiết. Xu hướng này phản ánh nhu cầu ngày càng tăng về trí tuệ nhân tạo xanh và bền vững trong tương lai.

2.5. Thảo luận cá nhân và triển vọng

2.5.1. Mối quan hệ đồng tiến hóa giữa kiến trúc phần cứng và mô hình AI

Sự phát triển của các mô hình trí tuệ nhân tạo quy mô lớn và kiến trúc phần cứng không diễn ra một cách độc lập, mà tồn tại mối quan hệ đồng tiến hóa chặt chẽ. Khi các mô hình AI ngày càng mở rộng về quy mô và độ phức tạp, nhu cầu về năng lực tính toán, bộ nhớ và băng thông tăng lên nhanh chóng, thúc đẩy sự ra đời của các kiến trúc phần cứng mới.

Mối quan hệ này thể hiện rõ nét trong sự phổ biến của kiến trúc Transformer, vốn phù hợp với các nền tảng phần cứng có khả năng tính toán song song rất cao. Các đặc điểm của phần cứng, chẳng hạn như cấu trúc bộ nhớ và khả năng truyền thông giữa các thiết bị, ngày càng ảnh hưởng trực tiếp đến quyết định thiết kế mô hình và thuật toán huấn luyện. Điều này dẫn đến xu hướng đồng thiết kế phần cứng - thuật toán, trong đó cả hai yếu tố được phát triển song song nhằm tối ưu hóa hiệu năng tổng thể của hệ thống AI.

2.5.2. So sánh các xu hướng kiến trúc phần cứng hiện nay

Kiến trúc phần cứng cho AI quy mô lớn hiện nay có thể được phân loại thành ba xu hướng chính: đa dụng hóa hiệu năng cao, chuyên biệt hóa và hệ thống hóa ở quy mô lớn. GPU đại diện cho hướng tiếp cận đa dụng hóa, với khả năng xử lý song song mạnh mẽ và hệ sinh thái phần mềm phong phú, phù hợp cho nhiều loại mô hình và ứng dụng AI khác nhau. Trong khi đó, các bộ tăng tốc AI chuyên biệt như TPU và ASIC tập trung tối ưu cho một lớp tác vụ nhất định, mang lại hiệu năng trên mỗi watt cao hơn nhưng đánh đổi bằng tính linh hoạt.

Song song với xu hướng chuyên biệt hóa là sự phát triển của các hệ thống tính toán phân tán quy mô lớn, nơi hiệu năng không chỉ phụ thuộc vào từng bộ xử lý đơn lẻ mà còn vào kiến trúc liên kết và cơ chế truyền thông giữa các nút. Xu hướng lấy bộ nhớ làm trung tâm được chú trọng nhằm giải quyết nút thắt cổ chai về băng thông và độ trễ. Mỗi xu hướng đều có ưu điểm và hạn chế riêng và việc kết hợp hiệu quả các hướng tiếp cận này là chìa khóa để xây dựng các hệ thống AI hiệu năng cao.

2.5.3. Tác động lâu dài của xu hướng phần cứng đối với AI quy mô lớn

Về lâu dài, các xu hướng kiến trúc phần cứng hiện nay sẽ định hình cách thức phát triển và ứng dụng trí tuệ nhân tạo quy mô lớn. Sự chuyên biệt hóa phần cứng có thể dẫn đến sự phân mảnh hệ sinh thái, nơi các mô hình AI phải được thiết kế phù hợp với từng nền tảng phần cứng cụ thể. Điều này vừa mở ra cơ hội tối ưu hóa sâu, vừa đặt ra thách thức về khả năng tương thích và tái sử dụng mô hình. Bên cạnh đó, việc tập trung vào hiệu quả năng lượng và phát triển bền vững sẽ ngày càng trở thành yếu tố quyết định trong nghiên cứu và triển khai AI quy mô lớn.

2.6. Thách thức và định hướng nghiên cứu trong tương lai

2.6.1. Giới hạn về năng lượng và tài nguyên

Một trong những thách thức lớn nhất đối với sự phát triển của các mô hình trí tuệ nhân tạo quy mô lớn là giới hạn về năng lượng và tài nguyên tính toán. Khi kích thước mô hình và tập dữ liệu huấn luyện tiếp tục tăng, nhu cầu về năng lực tính toán, bộ nhớ và băng thông truyền dữ liệu tăng theo cấp số nhân. Điều này không chỉ làm gia tăng chi phí đầu tư và vận hành hạ tầng phần cứng, mà còn đặt ra những rào cản về khả năng tiếp cận công nghệ đối với các tổ chức nghiên cứu có nguồn lực hạn chế. Hơn nữa, các giới hạn vật lý của công nghệ bán dẫn, những khó khăn trong việc tiếp tục thu nhỏ kích thước transistor, đang làm suy giảm tốc độ cải thiện hiệu năng phần cứng theo thời gian. Những giới hạn này buộc cộng đồng nghiên cứu phải tìm kiếm các giải pháp mới nhằm sử dụng hiệu quả hơn các tài nguyên phần cứng hiện có, thay vì chỉ dựa vào việc mở rộng quy mô phần cứng truyền thống.

2.6.2. Vấn đề bền vững và trí tuệ nhân tạo xanh

Việc huấn luyện các mô hình lớn tiêu thụ lượng điện năng đáng kể, kéo theo lượng phát thải carbon cao, đặc biệt khi sử dụng các trung tâm dữ liệu quy mô lớn. Do đó, vấn đề phát triển trí tuệ nhân tạo xanh đang trở thành một hướng nghiên cứu quan

trọng trong cả lĩnh vực phần cứng và thuật toán. Từ góc độ phần cứng AI, các nghiên cứu trong tương lai cần tập trung vào việc nâng cao hiệu năng, giảm tiêu thụ năng lượng và tối ưu hóa hiệu suất hệ thống tổng thể. Kết hợp giải pháp giữa phần cứng và phần mềm, chẳng hạn như nén mô hình, lượng tử hóa và điều chỉnh độ chính xác tính toán, có thể góp phần giảm chi phí năng lượng mà không làm suy giảm chất lượng mô hình. Những nỗ lực này đóng vai trò then chốt trong việc đảm bảo sự phát triển bền vững của trí tuệ nhân tạo.

2.6.3. Định hướng nghiên cứu kiến trúc phần cứng cho AI thế hệ tiếp theo

Trong bối cảnh hiện tại, nghiên cứu kiến trúc phần cứng cho AI thế hệ tiếp theo cần hướng tới các giải pháp đột phá, vượt ra ngoài các kiến trúc truyền thống. Một hướng nghiên cứu tiềm năng là phát triển các kiến trúc phần cứng siêu chuyên biệt, được đồng thiết kế chặt chẽ với các mô hình AI, nhằm tối ưu hóa toàn diện cả về hiệu năng, năng lượng và khả năng mở rộng.

Ngoài ra, các kiến trúc mới như tính toán gần bộ nhớ, tính toán trong bộ nhớ và các hệ thống kết hợp nhiều loại bộ xử lý khác nhau đang mở ra những khả năng mới cho AI quy mô lớn. Việc tận dụng các công nghệ bán dẫn tiên tiến và các mô hình tính toán phi truyền thống, có thể giúp vượt qua những giới hạn hiện tại về năng lượng và hiệu năng. Trong tương lai, sự kết hợp giữa kiến trúc phần cứng tiên tiến, thuật toán tối ưu và tư duy phát triển bền vững sẽ đóng vai trò quyết định trong việc định hình thế hệ tiếp theo của trí tuệ nhân tạo.

3. Kết luận

Thông qua việc phân tích đặc điểm của các mô hình AI hiện đại, các kiến trúc phần cứng phục vụ trí tuệ nhân tạo, cũng như tác động của phần cứng đến hiệu năng, khả năng mở rộng và hiệu quả năng lượng, nghiên cứu đã làm rõ mối quan hệ mật thiết giữa phần cứng và sự tiến bộ của trí tuệ nhân tạo. Bên cạnh đó, bài báo cũng tổng hợp các xu hướng kiến trúc phần cứng nổi bật hiện nay, bao gồm chuyên biệt hóa bộ tăng tốc, đồng thiết kế phần cứng - thuật toán, kiến trúc lấy bộ nhớ làm trung tâm và hệ thống tính toán phân tán quy mô lớn. Phần cứng không chỉ là nền tảng thực thi mà còn ảnh hưởng trực tiếp đến thiết kế mô hình, chiến lược huấn luyện và khả năng triển khai trong thực tiễn. Sự đồng tiến hóa giữa kiến trúc phần cứng và mô hình AI đang trở thành yếu tố then chốt, cho phép mở rộng quy mô mô hình, nâng cao hiệu năng. Trên cơ sở phân tích và thảo luận, tác giả đề xuất một số hướng nghiên cứu, bao gồm việc tiếp tục phát triển các kiến trúc phần cứng chuyên biệt cho AI, tăng cường nghiên cứu đồng thiết kế phần cứng - thuật toán và chú trọng đến các giải pháp hướng tới trí tuệ nhân tạo xanh và phát triển bền vững ■

Tài liệu tham khảo

[1]. M. D. Hill and D. A. Wood (2019). *The architectural implications of artificial intelligence.* IEEE Computer, vol. 52, no. 10, pp. 17 - 30, 2019.

[2]. J. Zhao et al (2024). *A survey of AI accelerators*. arXiv preprint arXiv:2401.04058.

[3]. A. Sharma et al (2024). *Hardware accelerators for artificial intelligence.* arXiv preprint arXiv:2411.13717.

[4]. J. Kaplan et al (2020). *Scaling laws for neural language models.* arXiv preprint arXiv:2001.08361, 2020.

[5]. D. Patterson et al (2019). *A new golden age for computer architecture.* Communications of the ACM, vol. 62, no. 2, pp. 48 - 60.

[6]. N. P. Jouppi et al (2018). *A Domain-Specific Architecture for Deep Neural Networks*. Communications of the ACM, vol. 61, no. 9, pp. 50 - 59.

[7]. D. Narayanan et al (2021). *Efficient Large-Scale Language Model Training on GPU Clusters.* Proc. MLSys.

[8]. Y. Kim et al (2019). *Revisiting Memory Bottlenecks in Large-Scale Deep Learning.* IEEE Micro, vol. 39, no. 5, pp. 64 - 75, 2019.

The trends and the impact of hardware architecture on the development of large-scale artificial intelligence (AI) models

Nguyen Van Thuan

Tien Giang University - Email: nguyenvanthuan@tgu.edu.vn.

Abstract: *This article studies the impact and trends of hardware on the development of large-scale artificial intelligence (AI) models in recent years. Training and deploying these models requires enormous computing power, far exceeding the capabilities of traditional hardware architectures. Therefore, hardware architecture has become a key factor, directly affecting the performance, scalability, cost, and energy efficiency of large-scale AI systems. The article also focuses on analyzing the role and influence of modern hardware architectures on the development of large-scale AI models, while synthesizing and evaluating prominent current hardware trends, including specialized AI accelerators, memory-centric architectures, and high-performance distributed computing systems. Based on this, the author discusses the prospects, remaining challenges, and proposes several research directions to promote the sustainable development of artificial intelligence in the future.*

Keywords: *Hardware architecture, large-scale artificial intelligence, AI accelerator, GPU, large-scale language model.*