
SNN

# VI MẠCH MẠNG NƠ-RON XUNG

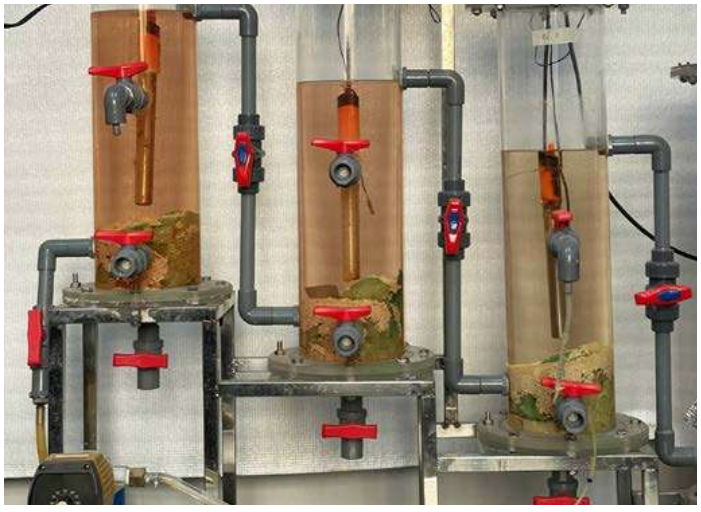
# XU HƯỚNG TẤT YẾU TRONG PHÁT TRIỂN CHIP AI HIỆN ĐẠI

ĐĂNG AN

**Xin GS cho biết xuất phát từ thực tiễn nào mà nhóm lựa chọn nghiên cứu về chip AI?**

Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANNs) ra đời đã thúc đẩy sự phát triển của các thuật toán học máy trong vài thập kỷ qua. Lấy cảm hứng từ bộ não sinh học, mạng ANNs được xây dựng từ các phần tử tính toán được gọi là nơ-ron. Các nơ-ron này nhận một tập hợp các đầu vào có trọng số từ các nơ-ron ở lớp trước, có giá trị kích hoạt liên tục và sử dụng các hàm kích hoạt phi tuyến tính có thể đạo hàm. Các nơ-ron được nhóm lại thành các lớp, và nhiều lớp được xếp chồng lên nhau để tạo ra một mạng nơ-ron rất sâu. Tính chất có thể đạo hàm của các hàm kích hoạt này cho phép sử dụng các phương pháp tối ưu hóa dựa trên gradient, chẳng hạn như lan truyền ngược (back propagation) để huấn luyện mạng; tức là tinh chỉnh các tham số của mạng sao cho phù hợp với tập hợp các đầu ra mong muốn. Nhờ những tiến bộ vượt bậc về khả năng tính toán với nền tảng GPU gần đây, kết hợp với sự sẵn có của các tập dữ liệu có nhãn lớn, việc huấn luyện các mạng rất sâu này trở nên khả thi. Lĩnh vực nghiên cứu này được gọi là học sâu (Deep Learning - DL), và các mạng với nhiều lớp nơ -ron được gọi là mạng nơ -ron sâu (Deep Neural Network - DNN). DNN đã được áp dụng thành công trong nhiều lĩnh vực, bao gồm nhận diện hình ảnh, phát hiện đối tượng, nhận dạng giọng nói, hoặc thậm chí chơi cờ vây.

Mặc dù DNN được xây dựng bắt chước bộ não con người, nhưng vẫn có một số khác biệt cơ bản giữa cách DNN xử lý thông tin và cách bộ não con người hoạt động. Sự khác biệt quan trọng nhất nằm ở cách biểu diễn và truyền tải thông tin giữa các phần tử tính toán. DNN biểu diễn đầu vào dưới dạng các giá trị kích hoạt liên tục và những giá trị này được truyền đi cũng như tích lũy dưới dạng các đầu vào có trọng số đến các nơ-ron tiếp theo. Ngược lại, nơ-ron sinh học giao tiếp với nhau thông qua chuỗi xung điện gọi là spike (tín hiệu xung). Mỗi cặp nơ-ron hình thành một kết nối gọi là khớp thần kinh (synapse). Các spike này xuất hiện rời rạc theo thời



TÍNH TOÁN MÔ PHỎNG THẦN KINH (HAY TÍNH TOÁN NEUROMORPHIC) ĐƯỢC CHO LÀ CÓ HIỆU SUẤT NĂNG LƯỢNG CAO HƠN NHIỀU SO VỚI KIẾN TRÚC MÁY TÍNH TRUYỀN THỐNG NHỜ VÀO BẢN CHẤT XỬ LÝ DỰA TRÊN SỰ KIỆN (EVENT-DRIVEN COMPUTING). ĐÂY CÓ THỂ LÀ HƯỚNG ĐI QUAN TRỌNG TRONG TƯƠNG LAI ĐỂ CẢI THIỆN HIỆU SUẤT XỬ LÝ CÁC THUẬT TOÁN TRÍ TUỆ NHÂN TẠO (AI), ĐẶC BIỆT LÀ VỚI CÁC HỆ THỐNG HỌC SÂU. NHẬN THẤY XU THẾ TẤT YẾU NÀY, NHÓM NGHIÊN CỨU HỆ THỐNG TÍCH HỢP THÔNG MINH (SISLAB) - ĐHQGHN ĐÃ BẮT ĐẦU NGHIÊN CỨU THIẾT KẾ PHẦN CỨNG TĂNG TỐC CHO CÁC MẠNG NƠ-RON NHÂN TẠO TỪ NĂM 2015 VÀ CÓ NHIỀU CÔNG TRÌNH KHOA HỌC CÔNG BỐ VỀ CÁC KẾT QUẢ NGHIÊN CỨU NÀY. VỚI MÔ HÌNH CHIP MẠNG NƠ-RON XUNG (SNN), NHÓM NGHIÊN CỨU ĐÃ TÍCH HỢP CÁC GIẢI PHÁP HIỆU QUẢ ĐỂ GIẢI QUYẾT NHỮNG THÁCH THỨC TỒN TẠI. VIỆC NHÓM ĐÃ THÀNH CÔNG PHÁT TRIỂN PHẦN CỨNG TĂNG TỐC CHO THUẬT TOÁN AI MANG LẠI NHIỀU LỢI ÍCH QUAN TRỌNG, ĐẶC BIỆT LÀ TRONG BỐI CẢNH AI NGÀY CÀNG YÊU CẦU HIỆU SUẤT CAO VÀ TIÊU TỐN NHIỀU TÀI NGUYÊN TÍNH TOÁN. VNU MEDIA ĐÃ CÓ CUỘC PHỎNG VẤN GS.TS. TRẦN XUÂN TÚ - VIỆN TRƯỞNG VIỆN CÔNG NGHỆ THÔNG TIN ĐHQGHN, TRƯỞNG NHÓM SISLAB ĐỂ HIỂU RÕ HƠN VỀ NHỮNG KẾT QUẢ BAN ĐẦU TRONG NGHIÊN CỨU VỀ CHIP AI.


gian, và thông tin có thể được biểu diễn thông qua thời gian phát xung (spike timing) hoặc tần suất phát xung (spike rate) trong một khoảng thời gian nhất định. Một khác biệt quan trọng nữa là cách quá trình học diễn ra trong não bộ sinh học so với quá trình huấn luyện của DNN. Phương pháp học dựa trên gradient của DNN không phù hợp với cơ chế sinh học, bởi vì quá trình điều chỉnh cường độ kết nối giữa các nơ-ron trong bộ não phụ thuộc vào thời gian tương đối giữa các xung đầu vào và đầu ra, thay vì phụ thuộc vào toàn bộ mạng như trong lan truyền ngược. Thông tin cần thiết cho các quy tắc học tập này chỉ tồn tại cục bộ giữa từng cặp nơ-ron kết nối, chứ không liên quan đến các nơ-ron khác trong mạng. Đấy là lý do tại sao bộ não của con người tiêu thụ công suất rất nhỏ, chỉ khoảng 20 W trong khi một hệ thống DNN lại tiêu thụ một công suất rất lớn.

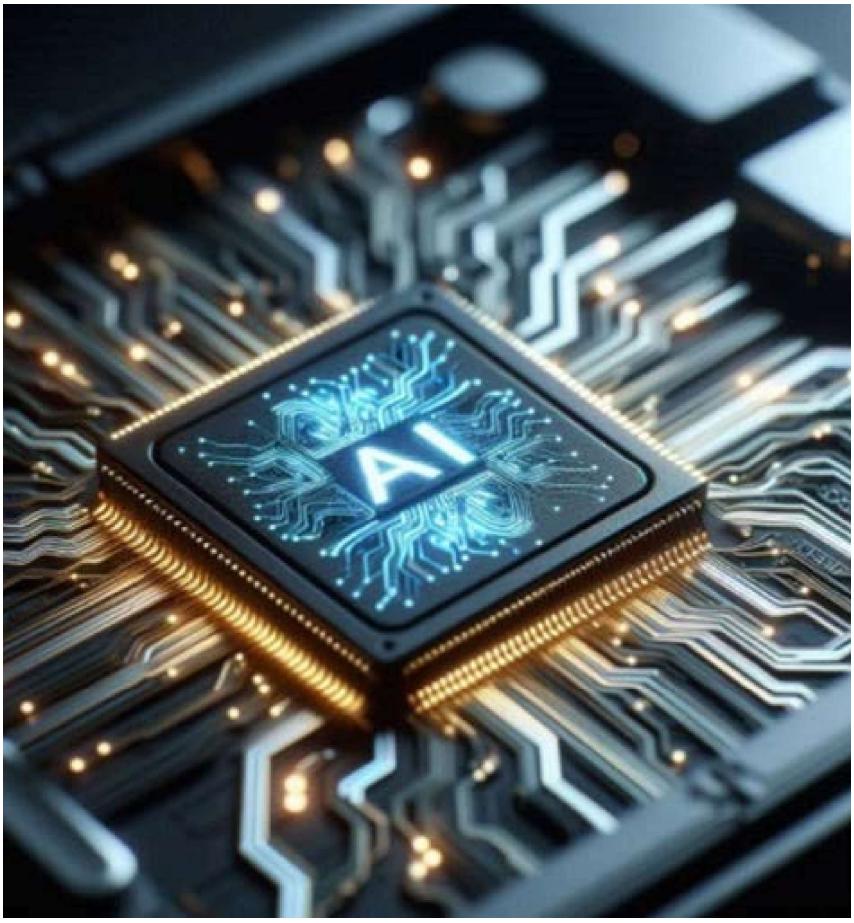
Những quan sát trên đã dẫn đến sự ra đời của mạng nơ-ron xung (Spiking Neural Network - SNN), được xem là thế hệ thứ ba của mạng nơ-ron nhân tạo. SNN có cơ chế hoạt động gần với hoạt động của bộ não con người hơn, bởi vì: (i) Các nơ-ron trong SNN cũng giao tiếp thông qua các xung điện (spikes); (ii) Các kết nối có trọng số giữa các cặp nơ-ron có thể được điều chỉnh thông qua quy luật huấn luyện dựa trên thời gian phát xung (Spike Timing Dependent Plasticity - STDP). So với DNN, SNN có một số lợi thế đáng chú ý. Thứ nhất là hiệu suất tính toán vì khả năng tính toán của SNN tương đương với ANN nhưng sử dụng ít phần tử tính toán hơn. Thứ hai là tiết kiệm năng lượng. Do các xung điện xuất hiện rời rạc theo thời gian và việc truyền thông tin chỉ diễn ra khi có sự kiện (event-driven), SNN giúp giảm tiêu thụ năng lượng so với DNN. Thứ ba là mô hình hóa dữ liệu theo thời gian. Thời điểm xảy ra xung đóng vai trò quan trọng trong các chiến lược mã hóa dữ liệu khác nhau. Với những ưu điểm này, SNN đã được áp dụng trong nhiều lĩnh vực như xử lý thị giác, nhận dạng giọng nói, và chẩn đoán y khoa. Trong những năm gần đây, việc kết hợp cấu trúc nhiều lớp của ANN với cơ chế tăng vọt lấy cảm hứng từ sinh học đã được nghiên cứu sâu rộng, dẫn đến sự ra đời của Mạng nơ-ron xung sâu (Deep Spiking Neural Networks- DSNN).

Khi kích thước mạng của DNN và SNN ngày càng tăng, độ phức tạp tính

toán của chúng cũng tăng theo, khiến cho việc thực thi trên kiến trúc máy tính Von Neumann truyền thống trở nên tốn thời gian và kém hiệu quả về năng lượng. Cộng đồng nghiên cứu thiết kế chip bán dẫn có độ tích hợp rất cao (VLSI) đã có nhiều nỗ lực trong việc phát triển kiến trúc phần cứng chuyên dụng để tăng tốc việc thực thi các thuật toán DNN và SNN. Do các thuật toán này mô phỏng quá trình tính toán của bộ não, việc thiết kế phần cứng cũng cần được lấy cảm hứng từ cấu trúc não bộ. Những hệ thống như vậy được gọi là hệ thống Tính toán mô phỏng thần kinh (Neuromorphic Computing). Tính toán Neuromorphic được cho là có hiệu suất năng lượng cao hơn nhiều so với kiến trúc máy tính truyền thống, nhờ vào bản chất xử lý dựa trên sự kiện (event-driven computing). Đây có thể là hướng đi quan trọng trong tương lai để cải thiện hiệu suất xử lý AI, đặc biệt là với các hệ thống học sâu và mạng nơ-ron xung.

Nhận thấy xu thế tất yếu này, Nhóm nghiên cứu SISLAB đã bắt đầu nghiên cứu thiết kế phần cứng tăng tốc cho các mạng nơ-ron nhân tạo ANN, CNN từ năm 2015 và có nhiều công trình khoa học công bố về các kết quả nghiên cứu này. Đến năm 2019, nhóm nghiên cứu SISLAB đã bắt đầu triển khai nghiên cứu thiết kế chip SNN (thế hệ thứ ba của mạng nơ-ron nhân tạo). Tuy nhiên, sự phát triển của các nền tảng tính toán thần kinh học (neuromorphic computing) hiệu quả vẫn đang phải đối mặt với những thách thức như:

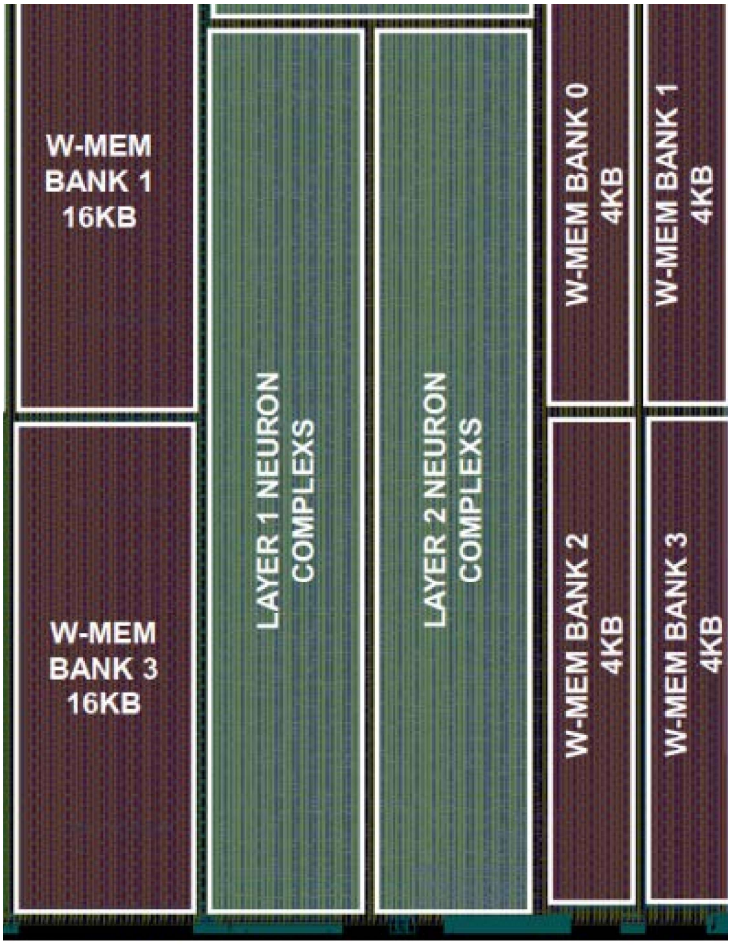
Thứ nhất, hiệu suất phần cứng chưa tối ưu cho SNNs trên các nền tảng nhúng. Kiến trúc phần cứng tiên tiến nhất hiện nay dành cho Mạng Nơ-ron xung (SNNs) trên nền tảng nhúng vẫn chưa đạt hiệu suất cao về chi phí phần cứng, dung lượng bộ nhớ và mức tiêu thụ năng lượng. Thành phần xử lý cơ bản trong bất kỳ triển khai SNN nào chính là nơ-ron. Do đó, để đảm bảo hệ thống hoạt động hiệu quả, cần phải có một kiến trúc phần cứng tối ưu cho nơ-ron. Khi kích thước mạng SNN mở rộng, yêu cầu về bộ nhớ cho các tham số mạng cũng tăng theo. Phần lớn các tham số mạng này chính là các trọng số đã được huấn luyện.



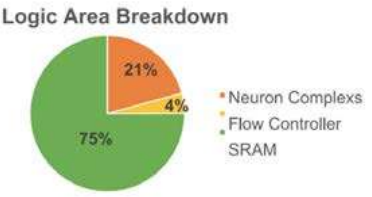

Thứ hai, thách thức về quản lý bộ nhớ trong quá trình suy luận (inference). Trong quá trình suy luận, các trọng số đã huấn luyện thường được tải từ bộ nhớ ngoài (DRAM) vào bộ nhớ đệm trên chip (on-chip buffer). Các trọng số trong bộ nhớ đệm có thể được tái sử dụng trong nhiều bước suy luận khác nhau. Nếu kích thước của trọng số vượt quá dung lượng bộ nhớ trên chip, thì hệ thống sẽ phải liên tục di chuyển dữ liệu giữa bộ nhớ trên chip và bộ nhớ ngoài. Điều này gây ra chi phí cao cả về mức tiêu thụ năng lượng và hiệu suất hệ thống, do truy cập DRAM chậm và tiêu tốn nhiều năng lượng. Để đảm bảo hiệu suất năng lượng tốt hơn, cần tìm cách giảm yêu cầu bộ nhớ cho các mạng SNN lớn.

**Đâu là bước đột phát trong nghiên cứu phát triển mô hình chip AI của nhóm nghiên cứu, thưa GS?**

Với mô hình chip AI đề xuất, chúng tôi đã tích hợp các giải pháp hiệu quả để giải quyết những thách thức đã đề cập ở trên. Việc triển khai Mạng Nơ-ron xung (SNN) hiệu quả trên nền tảng nhúng yêu cầu một thiết kế mới cho nơ-ron cơ bản. Trước tiên, chúng tôi đề xuất một thiết kế số mới cho nơ-ron Tích lũy và Phát xung (Integrate-and-Fire - LIF). Thiết kế này cần đáp ứng tiêu chí tiết kiệm chi phí phần cứng, giúp hệ thống có thể mở rộng để triển khai trên các mạng lớn hơn. Để chứng minh hiệu quả ở cấp độ hệ thống, chúng tôi sử dụng nơ-ron này làm lõi để triển khai một mạng SNN nhỏ, cố định với ba lớp




- Technology TSMC 65nm
- Vdd = 1.2 V
- Power consumption: 86mW
- Frequency: 167 MHz
- Core Area: 0.96 $mm^2$
- Energy Efficiency: 74 nJ/prediction
- Maximum throughput: 1.207M predictions/second



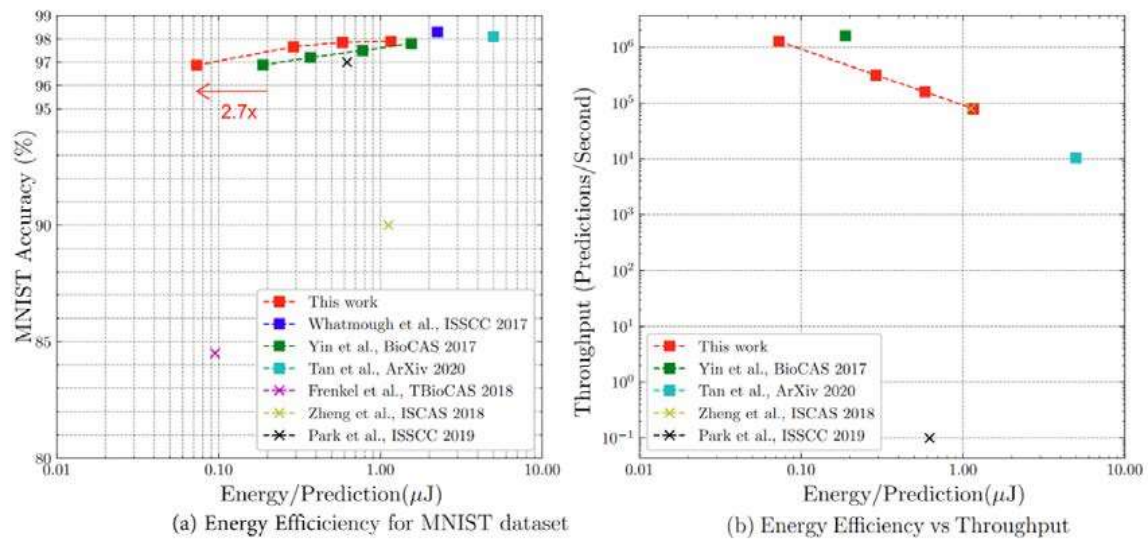

trên phần cứng. Nhằm cải thiện hơn nữa hiệu suất hệ thống, chúng tôi đề xuất một thuật toán mới cho SNN với trọng số ở định dạng tam phân (ternary format), giúp giảm đáng kể yêu cầu lưu trữ bộ nhớ.

Cụ thể, thiết kế số đơn giản cho nơ-ron LIF. Chúng tôi tập trung vào việc tối giản các tính năng bổ sung trong các thiết kế tiên tiến hiện nay, chỉ giữ lại cơ chế tích lũy và đặt lại trạng thái. Kết quả đạt được là một thiết kế gọn nhẹ hơn, giúp giảm chi phí diện tích phần cứng lên đến 3,2 lần. Để xác minh tính hiệu quả ở cấp độ hệ thống, chúng tôi cũng đề xuất một kiến trúc phần cứng mới để triển khai mạng SNN có 3 lớp kết nối đầy đủ, áp dụng cho bài toán nhận diện chữ số viết tay MNIST. Mạng SNN này được huấn luyện bằng phương pháp chuyển đổi từ ANN sang SNN (ANN-to-SNN conversion), với trọng số sử dụng định dạng số cố định 10-bit.

Bên cạnh đó, chúng tôi cũng đề xuất thuật toán huấn luyện mới cho SNN với trọng số tam phân (Ternary Weight Spiking Neural Networks - TW-SNN). Mục tiêu của thuật toán này là giảm yêu cầu bộ nhớ để lưu trữ các tham số mạng đã được huấn luyện. Để chứng minh hiệu suất tiết kiệm năng lượng của phương pháp, chúng tôi cũng đề xuất một kiến trúc phần cứng chuyên dụng cho TW-SNN.

Hệ thống phần cứng TW-SNN được đề xuất cho kiến trúc cố định 3 lớp được mô hình hóa bằng ngôn ngữ VHDL và triển khai trên công nghệ CMOS 65nm của hãng TSMC. Hệ thống bộ nhớ trọng số được tạo ra từ bộ biên dịch bộ nhớ và thiết kế được tổng hợp cũng như thực thi bằng các công cụ hỗ trợ thiết kế của hãng Synopsys.

Theo đó, tổng diện tích lõi sau bố trí (post-layout core area) là 0,96 $mm^2$, trong đó diện tích các cổng logic chiếm 0,24 $mm^2$ và bộ nhớ chiếm 0,72 $mm^2$. Hệ thống được kiểm tra với tập dữ liệu MNIST, sử dụng cấu hình mạng kết nối đầy đủ (fully connected network) với hai lớp ẩn, mỗi lớp gồm 256 nơ-ron. Với mức điện áp danh định 1,2V, thiết kế của chúng tôi đạt tần số mục tiêu 167 MHz và có mức tiêu thụ điện năng là 86 mW. Kết quả về hiệu suất năng lượng được xác định bằng công cụ Synopsys PrimeTime, sử dụng thông tin về hoạt động chuyển đổi dữ liệu (data switching activity) thu thập từ mô phỏng sau bố trí (post-layout simulation). Hình dưới đưa ra so sánh kết quả thiết kế của chúng tôi với các công trình khoa học khác cả về độ chính xác dự đoán cũng như năng lượng tiêu thụ và hiệu suất xử lý. Theo đó, độ chính xác của chip AI đề xuất có thể đạt tới 97 % - 98% trên tập dữ liệu MNIST trong khi công suất tiêu thụ giảm gần 3 lần so với các nghiên cứu gần đây.




**Xin GS cho biết giải pháp đưa lên chip AI có những ưu điểm gì?**

Việc phát triển phần cứng tăng tốc cho các thuật toán AI mang lại nhiều lợi ích quan trọng, đặc biệt là trong bối cảnh AI ngày càng yêu cầu hiệu suất cao và tiêu tốn nhiều tài nguyên tính toán. Một số lợi ích có thể kể đến như:

Thứ nhất, cải thiện tốc độ xử lý. Phần cứng chuyên dụng như GPU, TPU (Tensor Processing Unit), FPGA (Field-Programmable Gate Array), hay các chip Neuromorphic được thiết kế để xử lý khối lượng lớn các phép toán ma trận và tích chập nhanh hơn so với bộ vi xử lý thông thường. Tiếp đó, các bộ tăng tốc AI có khả năng thực hiện hàng nghìn đến hàng triệu phép tính đồng thời (xử lý song song) giúp giảm thời gian huấn luyện và suy luận của mô hình AI. Ví dụ, TPU của Google có thể tăng tốc xử lý mô hình TensorFlow lên hàng chục lần so với CPU thông thường.

Thứ hai, tiết kiệm năng lượng tiêu thụ và tăng hiệu suất. AI truyền thống trên CPU tiêu tốn nhiều điện năng do kiến trúc Von Neumann phải liên tục trao đổi dữ liệu giữa bộ nhớ và bộ vi xử lý. Phần cứng AI chuyên dụng được thiết kế để giảm độ trễ và mức tiêu thụ năng lượng, giúp tối ưu hóa hiệu suất trên mỗi watt điện. Ví dụ, TPU của Google có hiệu suất tính toán trên mỗi watt cao hơn 30-80 lần so với CPU thông thường. Chip Neuromorphic (chip tính toán thần kinh học) như chip Loihi của hãng Intel có thể mô phỏng mạng nơ-ron với mức tiêu thụ điện năng cực thấp, phù hợp cho AI nhúng hay AI trên điện toán biên.

Thứ ba, giảm chi phí vận hành AI quy mô lớn. Trong các hệ thống AI doanh nghiệp hoặc điện toán đám mây, việc tối ưu hóa phần cứng có thể giảm đáng kể chi phí vận hành. Các trung tâm dữ liệu AI sử dụng phần cứng chuyên dụng sẽ tối ưu chi phí so với việc sử dụng CPU đa năng. Ví dụ, Amazon AWS Inferentia là chip AI giúp giảm chi phí suy luận AI xuống còn 1/10 so với GPU thông thường.

Thứ tư là hỗ trợ triển khai AI trên thiết bị biên (Edge AI). AI không chỉ chạy trên đám mây mà còn cần chạy trên thiết bị biên như điện thoại, xe tự hành, camera thông minh. Chip AI nhúng giúp AI chạy mượt mà trên thiết bị nhỏ gọn mà không cần kết nối đến máy chủ. Ví dụ, chip Apple Neural Engine (ANE) trên iPhone giúp chạy AI trực tiếp trên điện thoại, tiết kiệm pin và tăng tốc độ xử lý ảnh. NVIDIA Jetson cung cấp AI nhúng cho robot và xe tự hành.

Thứ năm, mở ra khả năng ứng dụng AI trong thời gian thực. Với tốc độ cao và độ trễ thấp, phần cứng AI có thể hỗ trợ các ứng dụng AI thời gian thực như nhận diện khuôn mặt ngay lập tức trên camera an ninh, phân tích dữ liệu y tế để hỗ trợ bác sĩ ra quyết định nhanh hơn, hay AI trên xe tự lái có khả năng xử lý thông tin giao thông trong thời gian thực để đảm bảo an toàn. Ví dụ, hãng Tesla sử dụng chip tự thiết kế để xử lý AI trên xe tự lái, giúp giảm sự phụ thuộc vào đám mây và tăng tốc độ phản ứng.

**Theo GS, đâu là thách thức với nhóm nghiên cứu khi phải tối ưu hóa việc đưa mô hình AI lên phần cứng?**

Mặc dù có nhiều lợi ích như đề cập ở trên, việc triển khai cứng hóa các thuật toán AI hay nói cách khác là thiết kế chip AI gặp không ít khó khăn, thách thức.

Đầu tiên, khi các mô hình AI ngày

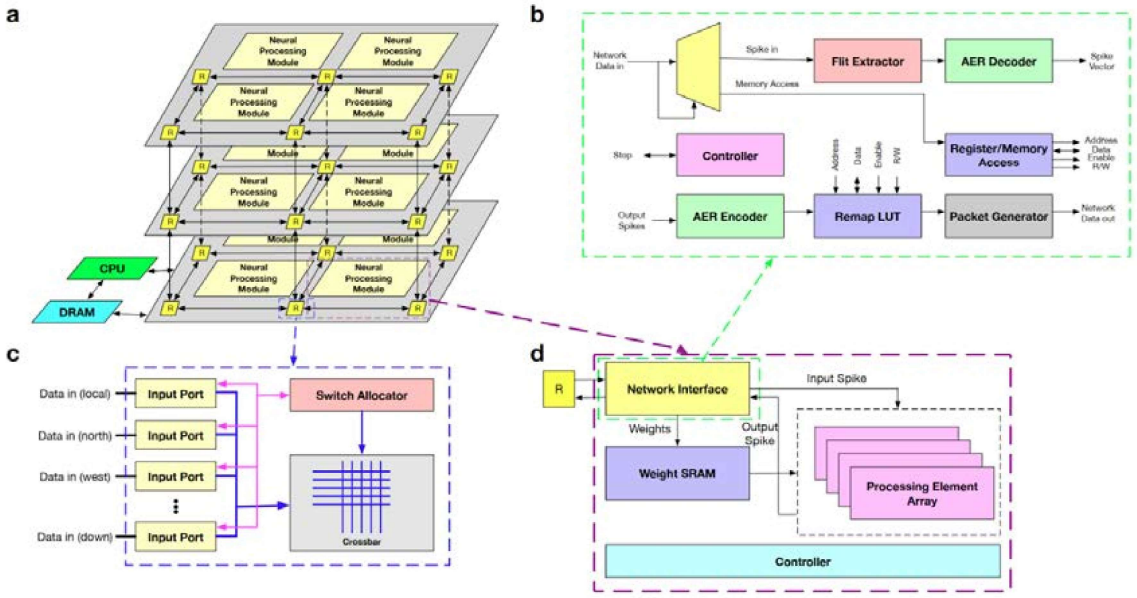

> Hiện tại nhóm nghiên cứu đang phát triển mô hình mạng trên chip 3 chiều (3D-NoC) tích hợp thiết kế phần cứng cho mạng SNN để có thể mở rộng cho các mô hình mạng SNN có kích thước lớn.

càng lớn, nhu cầu về khả năng truyền thông trên chip cũng trở thành một vấn đề nan giải. Các mô hình hiện đại như GPT-4 có hàng trăm tỷ tham số, đòi hỏi băng thông truyền dữ liệu cực cao giữa các phần tử tính toán. Nếu hệ thống truyền dữ liệu không được tối ưu, tình trạng tắc nghẽn (bottleneck) có thể xảy ra, làm giảm tốc độ suy luận và tăng mức tiêu thụ năng lượng. Các kiến trúc truyền thống như Von Neumann không phù hợp với AI do việc di chuyển dữ liệu liên tục giữa bộ nhớ và bộ xử lý gây ra độ trễ cao. Để khắc phục điều này, các kiến trúc như mạng trên chip (Network-on-Chip - NoC), bộ nhớ tốc độ cao như HBM (High Bandwidth Memory) và các giải pháp tính toán ngay trong bộ nhớ (In-Memory Computing) đang được nghiên cứu để giảm thiểu độ trễ và tối ưu luồng dữ liệu bên trong chip.

Một thách thức khác là khi cứng hóa thuật toán AI là khả năng mở rộng của phần cứng. Khi một con chip được sản xuất, nó có cấu trúc cố định và khó có thể thay đổi khi các thuật toán AI mới ra đời. Điều này dẫn đến nguy cơ lỗi thời nhanh chóng, đặc biệt khi AI phát triển theo từng tháng, trong khi một con chip AI có thể mất từ 18 tháng đến 24 tháng để thiết kế và sản xuất. Nếu phần cứng được thiết kế chỉ để tối ưu một thuật toán cụ thể, nó có thể không còn phù hợp khi có các mô hình AI tiên tiến hơn xuất hiện. Để giải quyết vấn đề này, các kiến trúc phần cứng có thể lập trình lại (Reconfigurable AI Accelerators) như FPGA hoặc các bộ xử lý RISC-V kết hợp các lõi AI đang được phát triển để giúp chip AI có thể thích ứng với nhiều mô hình khác nhau. Ngoài ra, xu hướng thiết kế hệ thống dựa trên công nghệ chiplet design cũng đang được áp dụng để tăng khả năng mở rộng mà không cần thiết kế lại toàn bộ hệ thống.

Bên cạnh đó, độ phức tạp trong thiết kế và sản xuất chip AI cũng là một thách thức lớn. Một hệ thống AI tích hợp đòi hỏi sự kết hợp của nhiều thành phần như bộ xử lý tensor (TPU), bộ nhớ tốc độ cao, mạng trên chip (NoC) và các bộ gia tốc chuyên biệt (NPU, DSP). Việc tối ưu tất cả các thành phần này trên một con chip đòi hỏi một quy trình thiết kế cực kỳ phức tạp, yêu cầu sự phối hợp chặt chẽ giữa nhóm phát triển phần cứng, phần mềm và thuật toán AI. Thêm vào đó, chi phí nghiên cứu và phát triển (R&D) cho một con chip AI có thể rất cao, và thời gian phát triển kéo dài cũng là một trở ngại lớn.

Tóm lại, việc thiết kế chip AI phải đối mặt với ba thách thức lớn: tối ưu truyền thông trên chip để xử lý mô hình lớn, giải quyết bài toán mở rộng của phần cứng để tránh lỗi thời nhanh chóng, và đối phó với độ phức tạp cao trong việc tích hợp các hệ thống AI vào một kiến trúc phần cứng hiệu quả.




**Vậy Nhóm nhiên cứu đã có giải pháp gì để giải quyết những thách thức này?**

Để vượt qua những rào cản này, các nhóm nghiên cứu cần tập trung vào phát triển công nghệ truyền thông trên chip tiên tiến, kiến trúc chip linh hoạt hơn, và phương pháp thiết kế hệ thống tối ưu từ cả góc độ phần cứng lẫn phần mềm.

**Khả năng ứng dụng trong thực tiễn của nghiên cứu như thế nào, thưa GS?**

Hiện nay các ứng dụng của trí tuệ nhân tạo và học sâu đang được ứng dụng rất rộng rãi vào trong các lĩnh vực của đời sống. Tuy nhiên, các ứng dụng của học sâu hiện nay đang gặp nhiều khó khăn khi triển khai sang các ứng dụng ở các thiết bị biên do yêu cầu tiêu tốn nhiều năng lượng và nhiều bộ nhớ lưu trữ. Mạng SNN là một trong những giải pháp hiệu quả để triển khai các ứng dụng của trí tuệ nhân tạo cho các thiết bị biên.

**Xin GS cho biết những hướng nghiên cứu tiếp theo trong thời gian tới?**

Hiện tại nhóm nghiên cứu đang phát triển mô hình mạng trên chip 3 chiều (3D-NoC) tích hợp thiết kế phần cứng cho mạng SNN để có thể mở rộng cho các mô hình mạng SNN có kích thước lớn. Đồng thời, nhóm nghiên cứu cũng dự kiến phối hợp với các nhóm nghiên cứu khác để sử dụng các công nghệ bộ nhớ tiên tiến như memristor trong thiết kế phần cứng cho mạng SNN, giúp giải quyết các bài toán thách thức nêu trên. Một trong những đề xuất của nhóm nghiên cứu là mô hình chip AI tích hợp với mạng trên chip 3 chiều.

**ĐHQGHN đang tạo lập các cơ chế, chính sách và môi trường nghiên cứu trong lĩnh vực bán dẫn, GS có thể chia sẻ một vài suy nghĩ về việc phát triển những hướng nghiên cứu mũi nhọn này ở ĐHQGHN để góp phần triển khai Nghị quyết số 57 của Bộ Chính trị về đột phá phát triển khoa học, công nghệ, đổi mới sáng tạo và chuyển đổi số quốc gia?**

ĐHQGHN là một trong các cơ sở giáo dục đại học hàng đầu của cả nước, giữ vai trò tiên phong trong việc phát triển và định hướng nền giáo dục đại học Việt Nam. Gắn bó với ĐHQGHN gần ba thập kỷ, tôi luôn cảm nhận được sự tâm huyết của các thế hệ lãnh đạo, cũng như niềm đam mê nghiên cứu khoa học và khát vọng chinh phục những đỉnh cao của các nhà khoa học trong hoạt động nghiên cứu khoa học, phát triển công nghệ và đào tạo, bồi dưỡng nhân tài cho đất nước. Tôi tin rằng, sự ra đời của Nghị quyết số 57 của Bộ chính trị về đột phá phát triển khoa học, công nghệ, đổi mới sáng tạo và chuyển đổi số sẽ là động lực mạnh mẽ, tạo bước chuyển quan trọng giúp ĐHQGHN vươn tầm, sẵn sàng hội nhập và cạnh tranh cùng các đại học hàng đầu thế giới trong kỷ nguyên mới.

**Xin trân trọng cảm ơn GS!**