|  |  |
| --- | --- |
| ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |



# ĐỀ CƯƠNG CHI TIẾT

|  |  |
| --- | --- |
| **TÊN ĐỀ TÀI TIẾNG VIỆT: BỘ TĂNG TỐC NHÂN MA TRẬN THƯA** | |
| **TÊN ĐỀ TÀI TIẾNG ANH:** | |
| **Cán bộ hướng dẫn:**   1. ThS. Trần Đại Dương | |
| **Thời gian thực hiện:** Từ ngày………………..đến ngày…………… | |
| **Sinh viên thực hiện:**  Dương Anh Khôi - 22520696  Lâm Xuân Phước - 22521154 | |
| **Nội dung đề tài**  **Tổng quan đề tài**:  Hiện nay, AI phát triển mạnh mẽ, đặc biệt trong học sâu, thúc đẩy nhiều lĩnh vực như y tế, tài chính và giao thông. Tuy nhiên, các mô hình AI phức tạp đòi hỏi tính toán lớn, tiêu thụ nhiều năng lượng và kém hiệu quả trên phần cứng truyền thống khi xử lý ma trận thưa – dữ liệu phổ biến với nhiều phần tử bằng 0. Điều này gây chậm trễ và lãng phí tài nguyên, cản trở ứng dụng AI tiên tiến. Vì vậy, bộ tăng tốc nhân ma trận thưa ra đời để tối ưu hóa tính toán, tập trung vào phần tử khác 0. Công nghệ này giúp tăng tốc độ, giảm năng lượng, nâng cao hiệu suất, hỗ trợ sự đổi mới AI trong tương lai.  Năm 2011, nhà nghiên cứu Ngô Thị Nhạn đã công bố luận văn "Thiết kế nhân ma trận thưa với vector trong tính toán song song và ứng dụng" tập trung vào tối ưu hóa phép nhân ma trận thưa-vector (SpMV) bằng phương pháp song song. Nghiên cứu này mang lại những ưu điểm như đạt được hiệu suất tính toán cao nhờ thuật toán Mondrian phân hoạch 2D thông minh kết hợp với thuật toán Biên địa phương (LA), giúp cân bằng tải (ε=0.03) và giảm 30-50% lượng truyền thông so với phương pháp truyền thống. Tuy nhiên, nghiên cứu vẫn còn một số nhược điểm cần khắc phục như hiệu suất giảm 15-20% với các ma trận có cấu trúc không đều. Về mặt triển khai, độ phức tạp bộ nhớ O(nz+P²) gây khó khăn khi mở rộng quy mô. Ngoài ra, giải pháp mới chỉ dừng lại ở SpMV.  Năm 2021, Matthew Denton trong luận văn "Acceleration of Sparse Matrix Multiplication Using Bit-Serial Arithmetic" đã phát triển bộ tăng tốc nhân ma trận thưa (SpMV) trên FPGA bằng kiến trúc bit-serial, tập trung vào ứng dụng học sâu. Tác giả sử dụng phép cộng bit-serial và biểu diễn Canonical Signed Digit (CSD) để loại bỏ phép tính thừa với giá trị 0, đồng thời tối ưu hóa phần cứng thông qua phân phối dữ liệu song song và tận dụng tính chất cố định của ma trận trọng số. Ưu điểm của giải pháp này là đạt được độ trễ cực thấp dưới 120ns, nhanh hơn 50–86 lần so với GPU Volta V100 và 4–47 lần so với bộ tăng tốc SIGMA, đồng thời giảm 17% tài nguyên FPGA nhờ biểu diễn CSD. Thiết kế phù hợp cho các ứng dụng đòi hỏi thời gian phản hồi tức thời trong môi trường điện toán biên như hệ thống nhúng và IoT. Nhược điểm chính của thiết kế là hiện tại chỉ hỗ trợ ma trận trọng số cố định trong quá trình suy luận, hạn chế khả năng áp dụng cho các mô hình cần cập nhật động.  Năm 2024, nhóm tác giả Trần Như Ý, Huỳnh Trung Hiếu và Phạm Thế Bảo trong bài báo "An Overview of Parallel Computing Models for Image Segmentation with Neut" đã nghiên cứu việc tăng tốc thuật toán Neut (Normalized Cuts) cho phân đoạn ảnh bằng các mô hình song song trên GPU và CPU. Trọng tâm là tối ưu phép tính ma trận thưa-vector (SpMV) trong bài toán trị riêng của ma trận Laplace, vốn chiếm phần lớn thời gian xử lý. Kết quả thực nghiệm cho thấy GPU (sử dụng CUDA/CUBLAS) giảm thời gian tính toán xuống 2.8 giây cho ảnh 1024×1024 so với 70.4 giây trên CPU, nhờ tận dụng kiến trúc đa lõi và bộ nhớ chia sẻ. Tuy nhiên, nghiên cứu chưa đi sâu vào các kỹ thuật cụ thể để tối ưu định dạng lưu trữ ma trận thưa (CSR, COO) hay cân bằng tải giữa các luồng. Ưu điểm là duy trì chất lượng phân đoạn ảnh, nhưng hạn chế nằm ở độ chính xác giảm khi kích thước ảnh quá lớn do giới hạn bộ nhớ GPU. Công trình mở ra hướng ứng dụng song song hóa cho các bài toán đại số tuyến tính quy mô lớn.  **Mục tiêu của đề tài:**  Mục tiêu chính là nghiên cứu và phát triển một kiến trúc phần cứng tăng tốc hiệu quả cho phép nhân ma trận thưa (Sparse Matrix-Vector Multiplication – SpMV), tập trung vào ứng dụng kỹ thuật nén dữ liệu để giảm lưu lượng truy cập bộ nhớ, tăng hiệu suất xử lý và tiết kiệm điện năng cho các hệ thống nhúng. Đặc biệt, nhóm lựa chọn khảo sát và tích hợp phương pháp Run-Length Encoding (RLE) như một giải pháp nhẹ, hiệu quả và phù hợp với kiến trúc phần cứng.  Từ mục tiêu tổng quát, nhóm chia thành 5 mục tiêu cụ thể như sau:   1. Nghiên cứu tổng quan các phương pháp nén dữ liệu hiện nay như Huffman, LZW, và Run-Length Encoding (RLE), nhằm xác định đặc điểm phù hợp với bài toán nén chuỗi dữ liệu trong ma trận thưa, đặc biệt là khả năng hiện thực hóa bằng phần cứng. 2. Đánh giá chi tiết ưu và nhược điểm của từng phương pháp nén theo các tiêu chí: tỷ lệ nén, độ phức tạp thuật toán, khả năng hiện thực phần cứng, và tính tương thích với đặc trưng của dữ liệu ma trận thưa. 3. So sánh hiệu quả của kỹ thuật RLE với các định dạng lưu trữ ma trận thưa như CSR, CSC và COO, theo các tiêu chí: mức độ nén, số lượng truy cập bộ nhớ, độ phức tạp khi tích hợp vào pipeline nhân SpMV. 4. Thiết kế và hiện thực mô hình thử nghiệm của phương pháp RLE trên các tập dữ liệu ma trận thưa thực tế (ví dụ: từ thư viện SuiteSparse), từ đó đo đạc hiệu quả nén, chi phí bộ nhớ, độ trễ xử lý, và khả năng tích hợp vào hệ thống nhúng. 5. Đề xuất kiến trúc phần cứng tích hợp RLE phù hợp với môi trường thực tế (như trên FPGA), cùng với các kỹ thuật tối ưu hóa pipeline, truy cập bộ nhớ và tính toán song song, nhằm tối đa hoá hiệu quả xử lý trong ứng dụng SpMV trên hệ thống nhúng.   Đề tài này tập trung vào việc tích hợp kỹ thuật nén dữ liệu Run-length Encoding (RLE) vào quá trình tăng tốc phép nhân ma trận thưa (SpMV), với kỳ vọng giảm băng thông bộ nhớ và nâng cao hiệu suất phần cứng trên các nền tảng nhúng như FPGA hoặc GPU. Khác với các nghiên cứu trước đây vốn chủ yếu tập trung vào song song hóa thuật toán (Ngô Thị Nhạn), tối ưu bit-serial (Matthew Denton), hoặc tăng cường xử lý trên GPU (Trần Như Ý, Huỳnh Trung Hiếu, Phạm Thế Bảo), đề tài này đề xuất một hướng tiếp cận mới – kết hợp cơ chế nén dữ liệu ngay trong quá trình thiết kế phần cứng tăng tốc. Cách tiếp cận này được kỳ vọng sẽ tận dụng tốt hơn tài nguyên hệ thống, đặc biệt trong các môi trường có giới hạn về bộ nhớ và năng lượng. Ngoài ra, đề tài cũng hướng tới khả năng ứng dụng trong các mô hình học sâu thực tế, nhằm đánh giá mức độ phù hợp và mở rộng của phương pháp đề xuất.  **Phương pháp nghiên cứu**   1. Phương pháp nghiên cứu lý thuyết: Nhóm sẽ khảo sát và phân tích các định dạng lưu trữ ma trận thưa phổ biến như CSR, CSC và COO để làm rõ đặc trưng dữ liệu và yêu cầu xử lý. Trên cơ sở đó, nhóm tập trung nghiên cứu kỹ thuật nén Run-length Encoding (RLE), đánh giá tiềm năng áp dụng cho dữ liệu ma trận thưa, từ góc độ khả năng nén, độ phức tạp thuật toán và mức độ phù hợp khi hiện thực bằng phần cứng. Ngoài ra, nhóm sẽ hệ thống hóa các kiến trúc tăng tốc SpMV hiện có để xác định điểm mạnh – điểm yếu làm nền tảng cho đề xuất của nhóm. 2. Phương pháp thiết kế hệ thống: Dựa trên kết quả phân tích lý thuyết, nhóm sẽ đề xuất thiết kế một mô-đun mã hóa và giải mã RLE dành riêng cho dữ liệu ma trận thưa, hướng tới tính hiệu quả khi triển khai bằng phần cứng. Thiết kế hệ thống sẽ bao gồm mảng xử lý song song (Processing Element Array), hỗ trợ thao tác trên các định dạng dữ liệu đã nén. Nhóm cũng sẽ xây dựng sơ bộ bộ quản lý truy cập bộ nhớ nhằm định tuyến dữ liệu phù hợp giữa các khối xử lý. 3. Phương pháp mô phỏng và đánh giá: Hệ thống sẽ được mô phỏng bằng SystemVerilog để kiểm thử tính đúng đắn về mặt chức năng. Các chỉ số như độ trễ, thông lượng, và tài nguyên FPGA sử dụng sẽ được thu thập nhằm đánh giá hiệu quả sơ bộ của phương pháp đề xuất. Việc phân tích hiệu năng sẽ được thực hiện ở nhiều cấu hình ma trận đầu vào khác nhau nhằm đánh giá tính ổn định và khả năng thích nghi của hệ thống. 4. Phương pháp triển khai thực tế: Sau giai đoạn mô phỏng, thiết kế sẽ được tổng hợp và triển khai trên board FPGA, làm bước trung gian trước khi cân nhắc tích hợp vào các hệ thống nhúng AI. Thông qua việc kiểm thử trên phần cứng thực tế, nhóm sẽ tiếp tục điều chỉnh và tối ưu hóa hệ thống nhằm nâng cao hiệu quả và giảm chi phí tài nguyên.   **Các nội dung chính và giới hạn của đề tài**:   1. Nội dung thực hiện: Đề tài tập trung vào nghiên cứu và phát triển một kiến trúc phần cứng tăng tốc cho phép nhân ma trận thưa với vector (SpMV), với định hướng ứng dụng trên các hệ thống nhúng hiệu năng cao. Nhóm sẽ tiến hành khảo sát và phân tích các định dạng lưu trữ ma trận thưa phổ biến hiện nay như CSR (Compressed Sparse Row) và COO (Coordinate Format), qua đó làm rõ đặc trưng phân bố dữ liệu, phương thức truy xuất bộ nhớ và mức độ phù hợp của từng định dạng đối với xử lý song song. Trên cơ sở đó, đề tài sẽ xem xét việc áp dụng kỹ thuật nén dữ liệu Run-length Encoding (RLE) như một hướng tiếp cận tiềm năng nhằm tối ưu hóa lưu trữ và truyền dữ liệu trong quá trình thực hiện phép nhân SpMV. Việc tích hợp RLE được kỳ vọng sẽ mang lại lợi ích trong các trường hợp dữ liệu có tính chất lặp lại cao – một đặc điểm thường xuất hiện trong các bài toán học sâu hoặc mô phỏng khoa học sử dụng ma trận thưa. Bên cạnh đó, đề tài cũng định hướng mở rộng bằng cách đề xuất các cải tiến nhằm hỗ trợ thêm nhiều dạng ma trận thưa khác nhau, tối ưu hóa kiến trúc cho nền tảng ASIC, hoặc tích hợp vào các hệ thống nhúng có yêu cầu cao về tính thời gian thực và tiết kiệm năng lượng. 2. Giới hạn của đề tài: Phạm vi nghiên cứu của đề tài giới hạn trong bài toán nhân ma trận thưa với vector (SpMV) – đây là phép toán xuất hiện thường xuyên trong các ứng dụng học sâu và mô phỏng vật lý. Tuy nhiên, đề tài không bao quát các dạng ma trận phức tạp hơn như ma trận thưa đối xứng, ma trận có mẫu phân bố không đều hoặc ma trận siêu thưa với độ phân tán cao. 3. Phương pháp đánh giá hệ thống: Việc đánh giá hệ thống sẽ được tiến hành thông qua hai giai đoạn chính: mô phỏng phần mềm và triển khai thực tế trên FPGA. Ở giai đoạn đầu, nhóm sẽ sử dụng mô phỏng chức năng bằng ngôn ngữ SystemVerilog để kiểm tra độ chính xác của thiết kế, đảm bảo rằng quá trình xử lý dữ liệu đầu vào cho kết quả đầu ra đúng như kỳ vọng. Sau khi đảm bảo tính chính xác chức năng, nhóm sẽ đánh giá tốc độ xử lý của hệ thống thông qua các chỉ số định lượng như thông lượng (tính bằng GFLOPS) và độ trễ (thời gian xử lý cho một tác vụ hoàn chỉnh). Các chỉ số này sẽ được đo trực tiếp từ hệ thống sau khi tổng hợp và triển khai trên FPGA, nhằm phản ánh hiệu suất thực tế trong điều kiện phần cứng. Thông qua quá trình đánh giá này, đề tài không chỉ kiểm chứng tính khả thi của thiết kế mà còn làm rõ các điểm mạnh, điểm hạn chế của đề tài. | |
| **Kế hoạch thực hiện (Phân công rõ người thực hiện)**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Công việc** | **Tháng 1** | **Tháng 2** | **Tháng 3** | **Tháng 4** | **Tháng 5** | **Tháng 6** | | **Nghiên cứu lý thuyết** | **Khôi, Phước** | **Khôi, Phước** | **Khôi, Phước** | **Khôi, Phước** |  |  | | **Thiết kế phần mềm** |  |  |  |  |  |  | | **Thiết kế phần cứng** |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Công việc** | **Tháng 7** | **Tháng 8** | **Tháng 9** | **Tháng 10** | **Tháng 11** | **Tháng 12** | | **Kiểm tra, đánh giá thiết kế** |  |  |  |  |  |  | | **Hiện thực trên FPGA** |  |  |  |  |  |  | | **Tổng hợp, viết báo cáo** |  |  |  |  |  |  | | |
| **Xác nhận của CBHD**  (Ký tên và ghi rõ họ tên) | **TP. HCM, ngày….tháng …..năm…..**  **Sinh viên**  (Ký tên và ghi rõ họ tên) |