(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0037865**

(51)[2020.01] **H04W 4/00**; G06N 3/00; H04W 28/00; G06F 15/00; H04L 12/00

(13) **B**

(21) 1-2020-06632

(22) 16/11/2020

(45) 25/12/2023 429

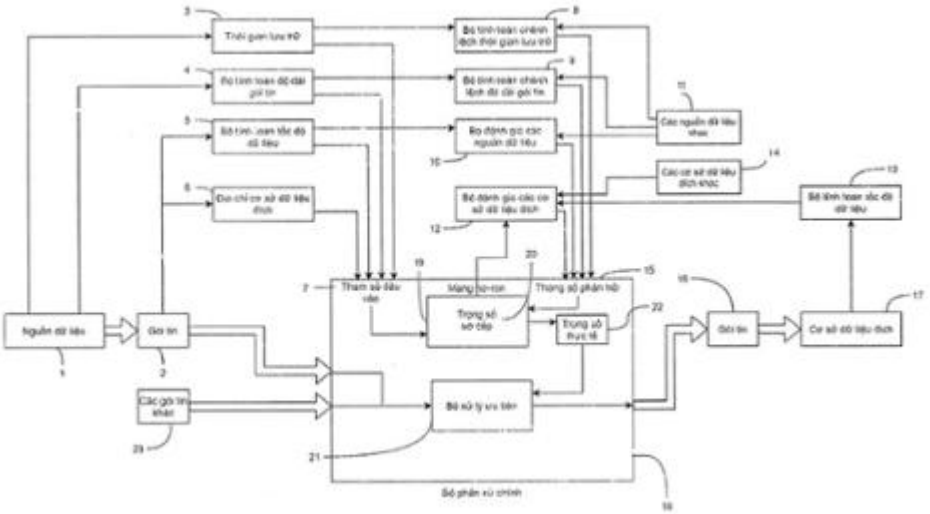
(43) 25/01/2021 394

(73) Trường Đại học Bách Khoa - Đại học Quốc Gia Thành phố Hồ Chí Minh (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh

(72) Hoàng Trang (VN).

(54) PHƯƠNG PHÁP QUẢN LÝ VÀ TĂNG TỐC ĐỘ TRUYỀN DỮ LIỆU QUA CỔNG GIAO TIẾP PCIE SỬ DỤNG CƠ CHẾ TRUY CẬP TRỰC TIẾP VÀ SỬ DỤNG MẠNG NƠRON ĐỂ ĐIỀU KHIỂN TRỌNG SỐ CỦA BỘ PHÂN XỬ DỮ LIỆU

(57) Sáng chế đề xuất phương pháp quản lý và tăng tốc độ truyền dữ liệu qua cổng giao tiếp PCIe sử dụng mạng nơron để điều khiển trọng số của bộ phân xử dữ liệu. Trong phương pháp theo sáng chế, khi phân xử việc truyền dữ liệu, một hệ thống các trọng số ưu tiên thứ tự truyền có thể thay đổi được, với hình thức học máy là nền tảng chính để quyết định mức độ ưu tiên của các loại dữ liệu khác nhau, trọng số càng lớn, đồng nghĩa với mức độ ưu tiên càng cao. Các trọng số này được gán riêng biệt cho từng gói tin, sẽ được quyết định dựa trên các tiêu chí: đặc điểm nguồn dữ liệu, đặc điểm cơ sở dữ liệu đích, chiều dài gói tin và thời gian lưu trữ của gói tin trong bộ nhớ. Giải pháp theo sáng chế giúp quản lý và tăng tốc độ truyền dữ liệu qua cổng giao tiếp PCIe một cách hiệu quả.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp thiết kế của một bộ quản lý các tác vụ truyền nhận và định địa chỉ gói tin trong giao tiếp giữa phần cứng mảng phần tử logic người dùng có thể lập trình được (FPGA - Field Programmable Gate Arrays) và máy tính chủ (Host) qua cổng giao tiếp PCI Express (PCIe), sử dụng cơ chế truy cập bộ nhớ trực tiếp (DMA - Direct Memory Access) giúp tăng tốc độ xử lý dữ liệu trong hệ thống và tăng truyền nhận dữ liệu qua cổng giao tiếp PCIe lên đến 8GB/s.

Tình trạng kỹ thuật của sáng chế

Cùng với sự phát triển của internet, kỹ thuật Internet kết nối vạn vật (IoT), hiện nay vấn đề tăng tốc độ truyền nhận dữ liệu giữa máy tính chủ và các thiết bị ngoại vi là vấn đề cực kỳ quan trọng, quyết định đến hiệu suất và chất lượng của cả hệ thống. Trên thị trường hiện nay có rất nhiều các giao tiếp (hay cổng kết nối) như: SATA, mSATA, M.2 hay PCI Express (viết tắt là PCIe), trong đó PCIe ra đời để đáp ứng yêu cầu truyền nhận dữ liệu ngày càng tăng theo sự phát triển của khoa học công nghệ.

Trong thiết bị định tuyến có bảo mật với chuẩn IPSec, để tăng tốc cho quá trình tính toán bảo mật và cung cấp khả năng cập nhật - tạo cấu hình lại yếu tố bảo mật, toàn bộ quá trình tính toán bảo mật theo chuẩn IPsec được đưa vào phần cứng FPGA (Mảng phần tử logic người dùng có thể lập trình được). Máy tính chủ (Host) sẽ đảm nhận các tác vụ mạng như định tuyến, phân tách mạng. Để đảm bảo tốc độ xử lý, dữ liệu được trao đổi giữa chip vi xử lý và FPGA thông qua giao tiếp tốc độ cao PCIe.

Trong vài năm qua, hiệu năng của giao tiếp Host và các thiết bị ngoại vi không ngừng tăng lên. Tốc độ giao tiếp PATA (Parallel ATA) tăng từ 33 MB/s lên 100 MB/s, trong khi hiệu năng của kết nối SATA (Serial ATA) nhảy vọt từ 1,5 Gb/s lên 6 Gb/s. Hiện nay, một số ổ cứng SSD sử dụng giao tiếp PCIe Gen 2 x4 (tốc độ thế hệ thứ 2, với 4 làn truyền dữ liệu), hỗ trợ lên đến 20 Gb/s (mỗi hướng). Do giao tiếp PCIe có thể đọc

và ghi đồng thời (full duplex), còn SATA chỉ có thể đọc hoặc ghi cùng một lúc (half-duplex), PCIe có khả năng tăng gấp đôi tốc độ 20 Gb/s đối với cả hai quá trình đọc và ghi dữ liệu giúp nó nhanh hơn gần 7 lần so với SATA.

Trên thế giới những năm vừa qua cũng đã có một vài nghiên cứu về việc tăng tốc độ truyền gửi dữ liệu, đặc biệt trong thiết bị định tuyến như kiến trúc RIFFA (Reusable Integration Framework for FPGA Accelerators), hay việc sử dụng các dòng NetFPGA với cấu hình và chi phí rất cao của Xilinx. Tuy nhiên, các thiết kế trên có giá thành cao, thiết kế về phần cứng khá lớn và phức tạp, do vậy dẫn đến việc khó tích hợp các thiết kế vào trong các ứng dụng, các hệ thống mà chỉ cần tận dụng đặc tính tốc độ truyền nhận dữ liệu nhanh.

PCIe được sử dụng phổ biến hiện nay nhờ có đặc tính tốc độ truyền nhận dữ liệu nhanh, hỗ trợ cho giao tiếp giữa 2 khối FPGA và máy tính chủ nhanh hơn. Vấn đề kỹ thuật cần thiết trong thiết kế các hệ thống hiện nay là phương pháp quản lý, phân xử dữ liệu thông qua PCIe cho hiệu quả, tận dụng tối đa đặc tính tốc độ cao này của PCIe.

Khi truyền dữ liệu với lưu lượng thấp hoặc ở mức thông thường như truyền dữ liệu qua các chuẩn giao tiếp COM, USB, v.v. mức độ khác biệt về tốc độ truyền các gói tin dung lượng thấp và có mức độ tương đồng khác nhau sẽ không có khác biệt nhiều bởi sự thay đổi mức ưu tiên truyền các gói tin sẽ không ảnh hưởng đáng kể tới tốc độ truyền cả gói tin. Tuy nhiên, đối với yêu cầu tốc độ truyền nhận dữ liệu cao ngày càng tăng theo sự phát triển của khoa học công nghệ hiện nay, giao tiếp PCIe là một lựa chọn phù hợp. PCIe có thể đọc và ghi đồng thời (full duplex), trong khi SATA chỉ có thể đọc hoặc ghi cùng một lúc (half-duplex), PCIe có khả năng tăng gấp đôi tốc độ 20 Gb/s đối với tải làm việc hỗn hợp (đọc và ghi), giúp nó nhanh hơn gần 7 lần so với SATA. Do đó, đối với giao tiếp PCIe khi truyền dữ liệu với tốc độ cao và dữ liệu có mức độ không tương đồng lớn, thì sự thay đổi mức ưu tiên truyền các gói tin sẽ tạo ra ảnh hưởng đáng kể. Hiện tại, theo hiểu biết của tác giả, chưa có giải pháp về một mô hình phân xử dữ liệu mà ở đó, một hệ thống các trọng số có thể thay đổi được, với hình thức học máy là nền tảng chính để quyết định mức độ ưu tiên của các loại dữ liệu khác nhau để áp dụng hiệu quả cho việc truyền dữ liệu trong các hệ thống phức tạp với việc truyền dữ liệu tốc độ lớn, xử lý dữ liệu liên tục và phức tạp khi sử dụng giao tiếp PCIe.

Đã có nhiều sáng chế cũng như các bài báo khoa học liên quan đến vấn đề nêu trên bằng cách thiết kế cơ chế phân xử dữ liệu giao tiếp giữa máy tính chủ và FPGA. Một trong những giải pháp liên quan gần nhất được thể hiện trong bài báo "An efficient weighted-round-robin algorithm for multiprocessor architectures", cũng như trong sáng chế US 7,512,148 B2 "Weighted round-robin arbitrator", hoặc trong sáng chế EP 1820109 B1 "Time-based weighted round robin arbiter", trong đó tác giả đề xuất một thiết kế phân xử dữ liệu mà ở đó các luồng dữ liệu không được phân xử như nhau mà có trọng số khác nhau. Trọng số này dựa trên đặc điểm của đích đến (ưu tiên đích đến càng cao thì trọng số càng cao để ưu tiên xử lý gói dữ liệu), hoặc dựa trên đặc điểm truyền trong hệ thống.

Với các cơ chế phân xử dữ liệu như trong các tài liệu này, hệ thống được thiết kế có thể giải quyết được các xung đột trong khi truyền/tiếp cận dữ liệu giữa nhiều hệ thống chủ động (master) và nhiều hệ thống thụ động (slave) hoặc nhiều bộ nhớ. Và từ đó, tốc độ lưu lượng của hệ thống được cải thiện đáng kể, có thể tăng thêm đến 41%.

Tuy nhiên, trong các mô hình của những sáng chế đã đề cập (US 7,512,148 B2 "Weighted round-robin arbitrator" và EP 1,820,109 B1 "Time-based weighted round robin arbiter"), việc các hệ số được xác định sẵn làm giảm hiệu quả cũng như quy mô ứng dụng của các mô hình này. Các gói tin có trọng số lớn nếu dùng cho những nguồn dữ liệu có lưu lượng nhỏ sẽ dẫn tới việc lãng phí tài nguyên, băng thông không được tận dụng triệt để; và ngược lại, các gói tin có trọng số nhỏ nếu dùng cho những nguồn dữ liệu có lưu lượng lớn sẽ khiến cho lượng lớn gói tin không được phân xử kịp thời, sau đó bị loại bỏ.

Do đó, hướng tiếp cận theo các mô hình này có nhược điểm là chỉ phù hợp trong các hệ thống đơn giản có sự truyền dữ liệu không nhiều, liên tục và phức tạp mà ở đó các cơ chế phân xử dữ liệu được đề cập ở trên với các trọng số được xác định trước. Trong các hệ thống phức tạp hơn với việc truyền, xử lý dữ liệu liên tục và phức tạp, cơ chế quản lý dữ liệu như vậy không tối ưu được hiệu năng của hệ thống.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là khắc phục nhược điểm của các giải pháp nêu trên và giải quyết các vấn đề kỹ thuật còn tồn tại.

Để đạt được mục đích trên, sáng chế này đề xuất phương pháp hệ thống và hệ thống quản lý và tăng tốc độ truyền dữ liệu qua cổng giao tiếp PCIe sử dụng mạng nơron để điều khiển trọng số của bộ phân xử dữ liệu mà ở đó, một hệ thống các trọng số có thể thay đổi được, với hình thức học máy là nền tảng chính để quyết định mức độ ưu tiên của các loại dữ liệu khác nhau. Các trọng số này, gán riêng biệt cho từng gói tin, sẽ được quyết định dựa trên 4 tiêu chí: đặc điểm nguồn dữ liệu, đặc điểm cơ sở dữ liệu đích, chiều dài gói tin và thời gian lưu trữ của gói tin trong bộ nhớ. Trọng số càng lớn, đồng nghĩa với mức độ ưu tiên càng cao.

Với giải pháp theo sáng chế, các nguồn dữ liệu khác nhau sẽ được gán với kết nối có trọng số phù hợp, cập nhật liên tục theo thời gian, qua đó tận dụng hiệu quả nguồn tài nguyên trong luân chuyển dữ liệu, đồng thời giảm thiểu việc mất mát thông tin.

Bên cạnh việc phân phối hiệu quả dữ liệu từ những nguồn khác nhau, giải pháp theo sáng chế cũng xem xét đến ảnh hưởng từ các cơ sở dữ liệu đích. Các trọng số của các gói tin sẽ được điều chỉnh tương ứng với tần suất, tốc độ xử lý dữ liệu của các cơ sở này. Các cơ sở dữ liệu thường xuyên nhận các gói dữ liệu, hay các cơ sở dữ liệu xử lý nhanh hơn các gói tin sẽ được ưu tiên trọng số cao; tài nguyên trong luân chuyển dữ liệu sẽ được dành nhiều hơn cho các cơ sở dữ liệu này.

Chiều dài của một gói tin cũng là một nhân tố khác có tác động đến việc điều chỉnh trọng số của gói tin đó. Gói tin càng dài, lượng dữ liệu mang theo càng lớn, do đó việc xử lý, luân chuyển sẽ mất nhiều thời gian hơn các gói tin có kích thước nhỏ hơn. Khi đó, việc gán trọng số lớn hơn cho chúng, như theo hệ thống theo sáng chế, là một giải pháp thích hợp: thông tin nhiều hơn sẽ được ưu tiên xử lý và di chuyển sớm hơn.

Một mặt khác, các gói tin trong nhiều trường hợp có thể bị lưu giữ tại bộ nhớ trong một thời gian dài vượt giới hạn cho phép, có thể là do trọng số của nó nhỏ hơn các gói tin khác đang được phân xử. Khi đó, các gói tin trên có khả năng bị kẹt lại trong bộ nhớ, không thể đến được cơ sở dữ liệu đích cho các quá trình xử lý tiếp theo. Mô hình phân xử dữ liệu trong sáng chế đưa ra cách giải quyết vấn đề trên bằng việc tính thêm thời gian lưu trữ của gói tin như một biến số trong giải thuật điều chỉnh hệ thống trọng số của mô hình.

Trong mô hình của sáng chế, như đã mô tả ở trên, các trọng số (gán cho từng gói tin) sẽ được quyết định dựa trên 4 tiêu chí, hay nói cách khác, dựa trên 4 tham số đầu vào. Mỗi gói tin sẽ được gán một tập các trọng số kể từ thời điểm nó bắt đầu đi vào mô hình phân xử. Kết quả của việc luân chuyển gói tin này, sẽ là cơ sở cho việc điều chỉnh lại trọng số để gói tin tiếp theo có thể được luân chuyển hiệu quả hơn. Sự điều chỉnh này, khiến cho các trọng số liên tục được cập nhật theo thời gian, theo tình hình thực tế của việc luân chuyển dữ liệu; chính là nền tảng của mô hình học máy trong sáng chế.

Với nhiều nét tương đồng với những hệ thống học máy hiện có, mô hình học máy trong sáng chế cũng được xây dựng dựa trên những yếu tố cơ bản sau:

- Đầu vào của máy học: các tham số đầu vào, đã mô tả ở trên, bao gồm tính chất nguồn dữ liệu, tính chất cơ sở dữ liệu đích, chiều dài gói tin và thời gian lưu trữ gói tin trong bộ nhớ.
- Thông số nội tại: các trọng số sơ cấp, là những vectơ bao gồm các phần tử là các thông số tương ứng với từng tham số đầu vào. Việc điều chỉnh trọng số của hệ thống theo sáng chế chính là việc điều chỉnh các vectơ trọng số sơ cấp này.
- Đầu ra của máy học: gồm các trọng số thực tế, là kết quả của quá trình tính toán dựa trên các tham số đầu vào cùng các trọng số sơ cấp; là thông số được sử dụng trực tiếp cho việc lựa chọn mức độ ưu tiên giữa các gói tin.
- Đánh giá, kiểm định: gồm các thông số đánh giá, phản hồi để quyết định mức độ phân xử hiệu quả của mô hình, bao gồm băng thông dữ liệu tại cơ sở dữ liệu đích, băng thông dữ liệu tại nguồn dữ liệu, mức độ chênh lệch về thời gian lưu trữ của các gói tin cùng nguồn, mức độ chênh lệch chiều dài dữ liệu giữa các gói tin của các nguồn dữ liệu khác nhau. Các thông số đánh giá này, là các thông tin phản hồi giúp mô hình máy học biết cách cập nhật lại các trọng số sơ cấp.

Theo một phương án thực hiện, sáng chế đề xuất hệ thống truyền dữ liệu sử dụng mạng nơron trong việc điều khiển trọng số của bộ phân xử dữ liệu qua giao tiếp PCIE để thực hiện phương pháp nêu trên, trong đó hệ thống này bao gồm các thành phần chính:

phần tính toán tham số đầu vào: bao gồm bộ ghi nhận thông tin thời gian lưu trữ của gói tin, bộ tính toán độ dài gói tin, bộ tính tốc độ dữ liệu của nguồn dữ liệu, và bộ tìm kiếm địa chỉ đích, các thông tin cần thiết cho tính toán được lấy từ nguồn dữ liệu, từ

gói tin, qua các bộ xử lý, ghi nhận, tính toán sẽ cung cấp cho bộ phân xử chính các tham số đầu vào;

phần đánh giá phản hồi bao gồm bộ tính toán chênh lệch thời gian lưu trữ gói tin, bộ tính toán chênh lệch độ dài gói tin, bộ đánh giá dữ liệu giữa các nguồn dữ liệu thu thập thông tin từ nguồn dữ liệu hiện tại cùng các nguồn dữ liệu khác, bộ tính toán tốc độ dữ liệu ở cơ sở dữ liệu đích lấy tốc độ của cơ sở dữ liệu đích hiện tại, bộ đánh giá dữ liệu giữa các cơ sở dữ liệu đích thu thập thông tin từ cơ sở dữ liệu đích hiện tại cùng các cơ sở dữ liệu đích khác, các thông tin đánh giá trên được thu thập lại và gửi đến bộ phân xử chính dưới dạng các thông số phản hồi;

bộ phân xử chính bao hàm một mạng nơron chứa đựng các trọng số sơ cấp cùng với bộ xử lý ưu tiên để quyết định thứ tự xử lý các gói tin, mạng nơron sẽ sử dụng các tham số đầu vào kết hợp với các trọng số sơ cấp để tính toán ra trọng số thực tế cung cấp cho bộ xử lý ưu tiên, mỗi khi kết thúc luân chuyển một gói tin, các thông số phản hồi sẽ được mạng nơron dùng để cập nhật lại các trọng số sơ cấp.

Nói chung, bản chất kỹ thuật của sáng chế này là việc sử dụng mô hình máy học để điều chỉnh liên tục, kịp thời các trọng số sơ cấp, kết hợp với các tham số đầu vào nhằm tạo ra các trọng số thực tế thích hợp cho từng gói tin, thông qua một quá trình phản hồi các đánh giá, kiểm định ứng với những gói tin đã luân chuyển trước đó.

Mô tả vắn tắt các hình vẽ

Hình 1 là hình vẽ minh họa tổng quát mô hình phân xử dữ liệu trong sáng chế.

Hình 2 là hình vẽ minh họa bộ phân xử chính, với nền tảng là mạng nơron dùng cho học máy.

Hình 3 là hình vẽ minh họa bộ tính toán tốc độ dữ liệu ở nguồn dữ liệu và cơ sở dữ liệu đích.

Hình 4 là hình vẽ minh họa bộ tính toán chênh lệch thời gian lưu trữ gói tin.

Hình 5 là hình vẽ minh họa bộ tính toán chênh lệch chiều dài gói tin cần xử lý tại nguồn dữ liệu.

Hình 6 là hình vẽ minh họa bộ đánh giá dữ liệu giữa các nguồn dữ liệu.

Hình 7 là hình vẽ minh họa bộ đánh giá dữ liệu giữa các cơ sở dữ liệu đích.

Hình 8 là hình vẽ minh họa bộ tính toán chiều dài gói tin.

Mô tả chi tiết sáng chế

Sau đây, giải pháp theo sáng chế sẽ được mô tả dựa vào các hình vẽ kèm theo. Trong đó, trên các hình vẽ, các số chỉ số giống nhau được dùng để chỉ dẫn các thành phần, chi tiết giống nhau hoặc tương tự nhau. Các hình vẽ chỉ nhằm mục đích giúp minh họa và hiểu rõ hơn giải pháp theo sáng chế, vì vậy, nguyên lý và phạm vi của sáng chế không giới hạn chỉ ở các hình vẽ này.

Hệ thống theo sáng chế, theo mô tả trong hình 1, vì mục đích tường minh, đơn giản hóa nên chỉ mô tả quá trình truyền gói tin thông qua 1 nguồn dữ liệu và một cơ sở dữ liệu đích duy nhất. Hệ thống đầy đủ thực tế sẽ bao gồm nhiều nguồn dữ liệu, cùng nhiều cơ sở dữ liệu đích khác nhau.

Để phân tích về mặt chức năng, hệ thống theo sáng chế được chia làm 3 phần chính:

- Phần tính toán tham số đầu vào: bao gồm bộ ghi nhận thông tin thời gian lưu trữ của gói tin 3, bộ tính toán độ dài gói tin 4, bộ tính tốc độ dữ liệu của nguồn dữ liệu 5, và bộ tìm kiếm địa chỉ đích 6. Các thông tin cần thiết cho tính toán được lấy từ nguồn dữ liệu 1, từ gói tin 2, qua các bộ xử lý, ghi nhận, tính toán sẽ cung cấp cho bộ phân xử chính 18 các tham số đầu vào 7.
- Phần đánh giá phản hồi như được thể hiện trên Hình 4: bao gồm bộ tính toán chênh lệch thời gian lưu trữ gói tin 8, bộ tính toán chênh lệch độ dài gói tin 9, bộ đánh giá dữ liệu giữa các nguồn dữ liệu 10 thu thập thông tin từ nguồn dữ liệu hiện tại 1 cùng các nguồn dữ liệu khác 11, bộ tính toán tốc độ dữ liệu ở cơ sở dữ liệu đích 13 lấy tốc độ của cơ sở dữ liệu đích hiện tại 17, bộ đánh giá dữ liệu giữa các cơ sở dữ liệu đích 12 thu thập thông tin từ cơ sở dữ liệu đích hiện tại 17 cùng các cơ sở dữ liệu đích khác 14. Các thông tin đánh giá trên được thu thập lại và gửi đến bộ phân xử chính 18 dưới dạng các thông số phản hồi 15.
- Bộ phân xử chính 18 bao hàm một mạng nơron 19 chứa đựng các trọng số sơ cấp 20 cùng với bộ xử lý ưu tiên 21 để quyết định thứ tự xử lý các gói tin. Mạng nơron 19 sẽ sử dụng các tham số đầu vào 7 kết hợp với các trọng số sơ cấp 20

để tính toán ra trọng số thực tế 22 cung cấp cho bộ xử lý ưu tiên 21. Mỗi khi kết thúc luân chuyển một gói tin, các thông số phản hồi 15 sẽ được mạng nơron 19 dùng để cập nhật lại các trọng số sơ cấp 20.

Quá trình hoạt động của mô hình, mô tả quá trình luân chuyển một gói tin, được cụ thể hóa bằng quy trình bên dưới và được minh họa trên cả hình 1, hình 2. Trong đó các bước được đánh thứ tự theo bảng chữ cái đi kèm với cùng một số thứ tự được hiểu là các bước thực hiện song song với nhau. Tuy nhiên, điều này không có nghĩa là các bước này phải hoàn thành cùng thời điểm với nhau. Thứ tự các bước chỉ là tượng trưng cho một gói tin cụ thể, trên thực tế chúng đều luôn hoạt động đồng thời:

- Bước 1: Đầu tiên, gói tin 2 được bắt đầu phát đi từ nguồn dữ liệu 1.
- Bước 2a: Bộ tính toán tốc độ dữ liệu 5 ứng với nguồn dữ liệu 1 trên so sánh mức chênh lệch thời gian đôi một giữa gói tin hiện tại 2 và các gói tin trước đó tại nguồn dữ liệu 1 này và tính toán được tốc độ dữ liệu của nguồn dữ liệu 1.
- Bước 2b: Nguồn dữ liệu 1 cung cấp thông tin về thời gian lưu trữ gói tin trong bộ nhớ 3 của gói tin 2.
- Bước 2c: Gói tin 2 được tính toán chiều dài bởi bộ tính toán chiều dài gói tin 4 (ứng với nguồn dữ liệu 1).
- Bước 2d: Địa chỉ cơ sở dữ liệu đích của gói tin 2 sẽ được trích xuất bởi bộ tìm kiếm địa chỉ đích 6 (ứng với nguồn dữ liệu 1).
- Bước 3: Các thông tin thu thập được (lấy mẫu đều đặn theo thời gian) sẽ được gửi đến bộ phân xử chính 18 dưới dạng các tham số đầu vào 7.
- Bước 4: Các thông số đầu vào 24 sẽ được qua mô hình mạng nơron 19 gồm nhiều lớp: bắt đầu từ lớp tích chập thứ nhất (gồm lớp tích chập 25 và hàm kích hoạt 26), qua nhiều lớp tích chập khác đến lớp tích chập cuối cùng (gồm lớp tích chập 27 và hàm kích hoạt 28) sau đó là một lớp kết nối đầy đủ 29 để tính được trọng số thực tế 31. Lớp tích chập và lớp kết nối đầy đủ sẽ sử dụng các thông số sơ cấp 30 để tính toán ra giá trị sử dụng cho các lớp tiếp theo.
- Bước 5: Thông số thực tế 31 (ứng với gói tin 2) sẽ được đưa đến bộ xử lý ưu tiên 32. Nếu trọng số của nó cao hơn tất cả các trọng số của các gói tin khác, gói tin 2 sẽ được ưu tiên gửi đi trước.
- Bước 6: Sau khi được gửi ra ngoài bộ phân xử chính 18 và trở thành gói tin mới 16, nó sẽ được chuyển tới cơ sở dữ liệu đích 17 của nó.

- Bước 7a: Bộ tính toán tốc độ dữ liệu 13 ứng với cơ sở dữ liệu đích 17 so sánh mức chênh lệch thời gian đôi một giữa gói tin hiện tại 16 và các gói tin trước đó nhận được tại cơ sở dữ liệu đích 17 này và tính toán được tốc độ dữ liệu của cơ sở dữ liệu đích 17.
- Bước 7b: Bộ đánh giá dữ liệu giữa các cơ sở dữ liệu đích 12 lấy thông tin từ bộ tính toán tốc độ dữ liệu 13 ứng với cơ sở dữ liệu đích 17, cũng như thông tin từ bộ tính toán tốc độ dữ liệu ứng với các cơ sở dữ liệu đích khác 14; từ đó đưa ra thông tin phản hồi về việc chỉnh sửa trọng số ứng với từng cơ sở dữ liệu đích.
- Bước 7c: Bộ đánh giá dữ liệu giữa các nguồn dữ liệu 10 lấy thông tin từ bộ tính toán tốc độ dữ liệu 5 ứng với nguồn dữ liệu 1, cũng như thông tin từ bộ tính toán tốc độ dữ liệu ứng với các nguồn dữ liệu khác 11; từ đó đưa ra thông tin phản hồi về việc chỉnh sửa trọng số ứng với từng nguồn dữ liệu.
- Bước 7d: Bộ tính toán chênh lệch thời gian lưu trữ gói tin 8 sẽ tính toán mức độ chênh lệch về thời gian lưu trữ gói tin của gói tin hiện tại 2 so với gói tin trước đó; sau đó so sánh mức chênh lệch này với những nguồn dữ liệu khác 11; từ đó đưa ra thông tin phản hồi về việc chỉnh sửa trọng số ứng với từng nguồn dữ liệu.
- Bước 7e: Bộ tính toán chênh lệch chiều dài gói tin 9 sẽ tính toán mức độ chênh lệch về chiều dài gói tin còn lại cần xử lý của gói tin hiện tại 2 (đã xử lý một phần chiều dài) so với gói tin trước đó; sau đó so sánh mức chênh lệch này với những nguồn dữ liệu khác 11; từ đó đưa ra thông tin phản hồi về việc chỉnh sửa trọng số ứng với từng nguồn dữ liệu.
- Bước 8: Các thông tin phản hồi thu thập được (lấy mẫu đều đặn theo thời gian) sẽ được gửi đến bộ phân xử chính 18 dưới dạng các thông số phản hồi 15.
- Bước 9: Các thông số phản hồi 33 được đưa vào các bộ lan truyền ngược đạo hàm: bắt đầu từ bộ lan truyền ngược của lớp kết nối đầy đủ 34, bộ lan truyền ngược của lớp tích chập cuối cùng 35 (theo chiều tính toán trọng số thực tế 31), qua nhiều bộ lan truyền ngược của các lớp tích chập khác, kết thúc ở bộ lan truyền ngược của lớp tích chập đầu tiên 36 (theo chiều tính toán trọng số thực tế 31). Ở từng bộ lan truyền ngược này, các thông số sơ cấp 30 ứng với từng lớp sẽ được cập nhật lại giá trị mới phù hợp với các thông tin phản hồi 33.

Các tham số đầu vào sẽ được lấy mẫu sau mỗi chu kì thời gian nhất định. Điều tương tự cũng diễn ra với các thông số phản hồi. Thời gian lấy mẫu không phụ thuộc vào hoạt

động của các khối khác ngoại trừ bộ phân xử chính 18. Nói cách khác, sau mỗi lần lấy mẫu, thông tin nào chưa được cập nhật mới thì nó sẽ được cập nhật trong lần lấy mẫu tiếp theo. Do đó thời gian lấy mẫu cần đủ dài để cập nhật thông tin hiệu quả, cũng như dành thời gian cho quá trình tính toán trọng số thực tế 22 và quá trình lan truyền ngược.

Chi tiết cụ thể về các khối chức năng con như sau:

- Bộ tính toán tốc độ dữ liệu ở nguồn dữ liệu và cơ sở dữ liệu đích như được thể hiện trên Hình 3: tốc độ luân chuyển dữ liệu của nguồn dữ liệu (hay cơ sở dữ liệu đích) sẽ tương ứng với tín hiệu clock của nguồn dữ liệu (hay cơ sở dữ liệu đích) đó 38. Tín hiệu xung nhịp (clock) cần đo 38 sẽ được đồng bộ hóa (bộ đồng bộ 39) với một tín hiệu clock mẫu 37. Bộ đếm 40 sẽ đếm số lần xuất hiện xung cạnh dương của tín hiệu nhịp đồng hồ (clock) 38 trong một khoảng thời gian định trước của tín hiệu clock mẫu 37. Số đếm này qua đó sẽ được tính toán ra tần số, hay tốc độ 41 của tín hiệu clock cần đo 38.
- Bộ tính toán chênh lệch thời gian lưu trữ gói tin như được thể hiện trên Hình 4: thông tin về thời gian lưu trữ gói tin hiện tại 42 sẽ được so sánh với thời gian lưu trữ gói tin trước đó 44 (đã ngăn miền thời gian bằng các flip-flop 43), sau đó đưa qua bộ trừ 45 (tính chênh lệch) để có được các tín hiệu dấu 46 và độ chênh lệch 47. Các tín hiệu này ứng với nguồn dữ liệu hiện tại sẽ được so sánh với các tín hiệu cùng loại của những nguồn dữ liệu khác 48 bởi bộ so sánh 49. Sau quá trình so sánh, bộ so sánh 49 sẽ quyết định mức tăng/giảm trọng số của từng nguồn dữ liệu và gửi thông tin này 50 đến bộ phân xử chính dưới dạng thông tin phản hồi 15.
- Bộ tính toán chênh lệch chiều dài gói tin như được thể hiện trên Hình 5: chiều dài gói tin hiện tại 51 (lấy từ bộ tính toán chiều dài gói tin 4) sẽ được so sánh với chiều dài gói tin trước đó 53 (đã ngăn miền thời gian bằng các flip-flop 52), sau đó đưa qua bộ trừ 54 (tính chênh lệch) để có được các tín hiệu dấu 55 và độ chênh lệch 56. Các tín hiệu này ứng với nguồn dữ liệu hiện tại sẽ được so sánh với các tín hiệu cùng loại của những nguồn dữ liệu khác 57 bởi bộ so sánh 58. Sau quá trình so sánh, bộ so sánh 58 sẽ quyết định mức tăng/giảm trọng số của từng nguồn dữ liệu và gửi thông tin này 59 đến bộ phân xử chính dưới dạng thông tin phản hồi 15.

- Bộ đánh giá dữ liệu giữa các nguồn dữ liệu như được thể hiện trên Hình 6: thông tin tốc độ nguồn dữ liệu (lấy từ bộ tính toán tốc độ dữ liệu 5) của các nguồn dữ liệu từ nguồn dữ liệu đầu tiên 60 đến nguồn dữ liệu cuối cùng 61 sẽ được đem so sánh bởi bộ so sánh 62. Sau quá trình so sánh, bộ so sánh 62 sẽ quyết định mức tăng/giảm trọng số của từng nguồn dữ liệu và gửi thông tin này 63 đến bộ phân xử chính dưới dạng thông tin phản hồi 15.
- Bộ đánh giá dữ liệu giữa các cơ sở dữ liệu đích như được thể hiện trên Hình 7: thông tin tốc độ cơ sở dữ liệu đích (lấy từ bộ tính toán tốc độ dữ liệu 13) của các cơ sở dữ liệu đích từ cơ sở dữ liệu đích đầu tiên 64 đến cơ sở dữ liệu đích cuối cùng 65 sẽ được đem so sánh bởi bộ so sánh 66. Sau quá trình so sánh, bộ so sánh 66 sẽ quyết định mức tăng/giảm trọng số của từng nguồn dữ liệu và gửi thông tin này 67 đến bộ phân xử chính dưới dạng thông tin phản hồi 15.
- Bộ tính toán chiều dài gói tin như được thể hiện trên Hình 8: Gói tin 68 sẽ được truyền nối tiếp từng 32 bit một bởi bộ truyền 32 bit 69. Bộ nhận diện bắt đầu/kết thúc gói tin 70 sẽ ghi nhận lại thời điểm nhận được thông tin bắt đầu/kết thúc và gửi chúng (các tín hiệu điều khiển 71, 72) đến bộ đếm 73 nhằm xác định được chiều dài gói tin 74.

Ví dụ thực hiện sáng chế

Sau đây là một ví dụ mô tả cách thức hoạt động của giải pháp hay mô hình phân xử dữ liệu trong sáng chế. Lưu ý nội dung ví dụ này chỉ cung cấp cho người đọc cái nhìn tổng quan về cách thức mà mô hình này hoạt động. Các thông số, quy mô, cấu trúc mô hình, đánh giá hiệu quả trong ví dụ này chỉ là tượng trưng và có ý nghĩa chỉ trong ví dụ này, khi áp dụng sáng chế này vào thực tế thì quy mô, cấu trúc mô hình, giá trị thông số, mức độ hiệu quả có thể sẽ không trùng khớp.

Như trình bày trong sáng chế, các tham số đầu vào của mô hình mà quyết định mức độ quan trọng của một gói tin, bao gồm tính chất nguồn dữ liệu, tính chất cơ sở dữ liệu đích, chiều dài gói tin và thời gian lưu trữ gói tin trong bộ nhớ.

Giả định có một bảng đánh giá mức độ quan trọng của từng tham số đầu vào trong những điều kiện khác nhau. Bảng này là do người dùng, người thiết kế định nghĩa, và là điểm chuẩn mà mô hình sẽ hướng tới việc học theo:

Thứ tự ưu tiên	Loại tham số	Điều kiện
1	Nguồn dữ liệu	Tốc độ dữ liệu lớn hơn 500Kbps
2	Chiều dài gói tin cần xử lý	Lớn hơn 1200 byte
3	Thời gian lưu trữ gói tin	Lớn hơn 100us
4	Nguồn dữ liệu	Tốc độ dữ liệu lớn hơn 200Kbps
5	Cơ sở dữ liệu đích	Tốc độ dữ liệu lớn hơn 250Kbps
6	Chiều dài gói tin cần xử lý	Lớn hơn 800 bytes

Khi đã có bảng dữ liệu này, người thiết kế sẽ căn chỉnh các hệ số nhân của các bộ tính toán chênh lệch thời gian lưu trữ, bộ tính toán chênh lệch chiều dài gói tin, bộ đánh giá các nguồn dữ liệu, bộ đánh giá các cơ sở dữ liệu đích. Các bộ tính toán nói trên có nhiệm vụ đánh giá các gói tin của cùng một loại tham số đầu vào, trong khi các hệ số nhân (hệ số dành cho thông số phản hồi) có tác dụng điều chỉnh chọn loại tham số đầu vào nào quan trọng hơn. Trong số các thông số phản hồi, bộ phân xử chính sẽ lấy thông số có giá trị cao nhất để làm mẫu thử cho quá trình lan truyền ngược của mạng nơron. Mức độ chênh lệch của mẫu thử so với trọng số thực tế cũng đồng nghĩa với mức độ điều chỉnh thông số nội tại của mạng nơron.

Giả định có gói tin X với các thông số đầu vào như sau:

Gói tin	Tốc độ dữ liệu của nguồn dữ liệu	Tốc độ dữ liệu của cơ sở dữ liệu đích	Chiều dài gói tin còn lại cần xử lý	Thời gian lưu trữ của gói tin
X	400Kbps	100Kbps	1400 byte	50us

Đồng thời giả định có bảng thông số nội tại:

Vị trí	Giá trị
W0 lớp 1	0,25
W1 lớp thứ 1	-0,3
W2 lớp thứ 1	0,6

W3 lớp thứ 1	0,1
W0 lớp thứ 2	-0.35
W1 lớp thứ 2	0,9

Các tham số đầu vào của gói tin X, sau khi được chuẩn hóa và biểu diễn dưới dạng một vectơ V, sẽ được sử dụng cùng các thông số nội tại (kí hiệu IW) để tính ra trọng số thực tế (kí hiệu RW).

$$RW = f(V, IW)$$

Giả sử kết quả của hàm trên là 3,09.

Thông số phản hồi thu được từ các bộ tính toán chênh lệch thời gian lưu trữ, bộ tính toán chênh lệch chiều dài gói tin, bộ đánh giá các nguồn dữ liệu, bộ đánh giá các cơ sở dữ liệu đích thu được lần lượt là 1,5, 2,67, 2,58, 4,99. Mẫu thử (kí hiệu EW) được lấy sẽ là 4,99. Mức độ chênh lệch của mẫu thử và RW sẽ là |4,99-3,09|=1,90. Giá trị 1,90 này sẽ thông qua quá trình lan truyền ngược để tính toán ra giá trị cần thiết để điều chỉnh các thông số nội tại của mạng nơron (kí hiệu DW).

$$DW = g(EW, RW)$$

$$IW = IW - DW$$

Nếu dùng các thông số nội tại mới được cập nhật để tính toán trọng số thực tế một lần nữa với gói tin A, giá trị trọng số thực tế sẽ tiến gần hơn về giá trị 4,99.

Như vậy, để mô hình hoạt động hiệu quả, mạng nơron cần trải qua một quá trình học nhằm điều chỉnh giá trị các thông số nội tại sao cho kết quả trọng số thực tế sẽ tuân theo bảng đánh giá mức độ quan trọng của người thiết kế.

Chú ý rằng mô hình không tác động đến thời gian luân chuyển 1 gói tin trong bộ phân xử mà nó chỉ quyết định mức độ ưu tiên của từng gói tin thông qua các tham số đầu vào. Đối với các gói tin có mức độ quan trọng, chẳng hạn như gói tin từ nguồn dữ liệu có tốc độ lớn hơn 1200 byte như trong ví dụ này, xác suất mà nó sẽ được ưu tiên nhanh hơn 5/6=83%, trong khi gói tin có mức độ ưu tiên sau cùng sẽ có xác suất được gửi đi đầu tiên chỉ là 1/6=17%. Nói cách khác, mô hình chỉ hoạt động hiệu quả nhất trong trường hợp có nhiều loại gói tin với mức độ ưu tiên khác nhau, các gói tin với mức

độ ưu tiên cao sẽ có hiệu suất luân chuyển cao hơn, trong khi ta phải chấp nhận các gói tin có mức độ ưu tiên thấp sẽ bị giảm hiệu suất lại; nếu dùng mô hình trong sáng chế cho những loại gói tin tương đồng thì ta chỉ thu được hiệu suất ngang bằng mô hình phân xử thông thường.

Yêu cầu bảo hộ

1. Phương pháp quản lý và tăng tốc độ truyền dữ liệu qua cổng giao tiếp PCIe sử dụng cơ chế truy cập trực tiếp và sử dụng mạng nơron để điều khiển trọng số của bộ phân xử dữ liệu, phương pháp này bao gồm các bước:

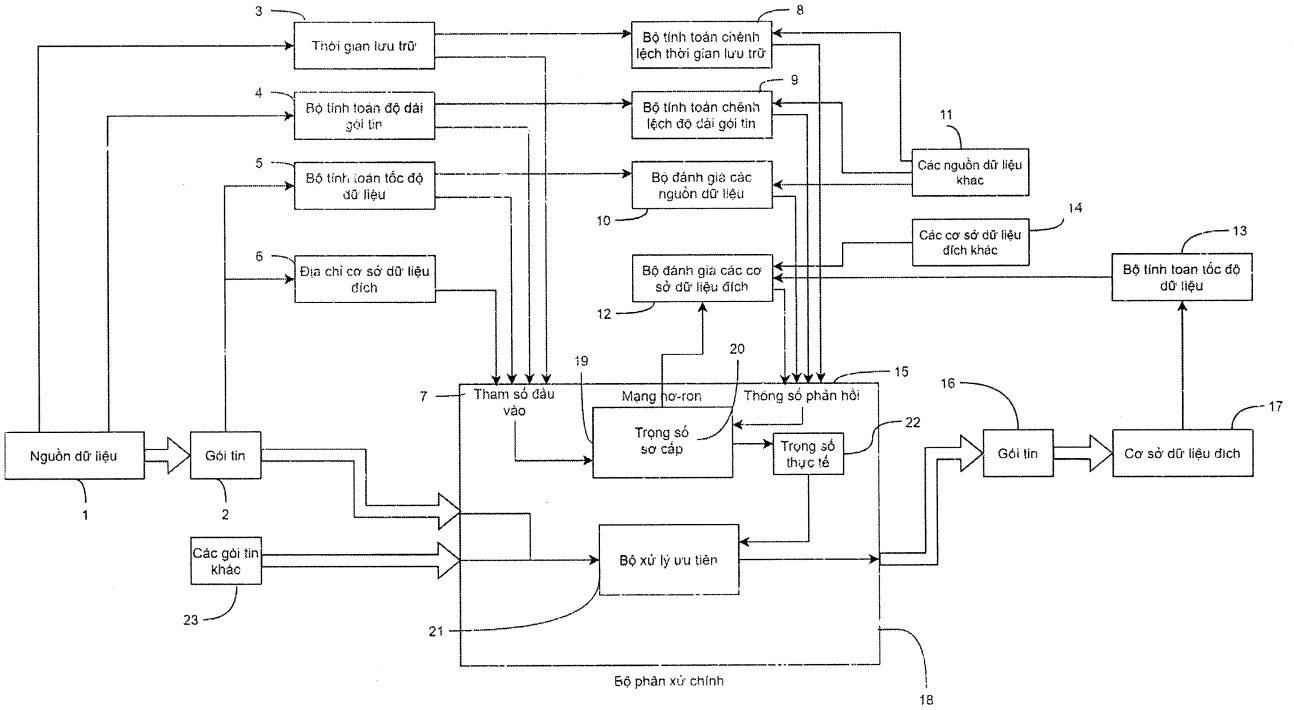
- bước 1: đầu tiên, gói tin (2) được bắt đầu phát đi từ nguồn dữ liệu (1);
- bước 2a: bộ tính toán tốc độ dữ liệu (5) ứng với nguồn dữ liệu (1) trên so sánh mức chênh lệch thời gian đôi một giữa gói tin hiện tại (2) và các gói tin trước đó tại nguồn dữ liệu (1) này và tính toán được tốc độ dữ liệu của nguồn dữ liệu (1);
- bước 2b: nguồn dữ liệu (1) cung cấp thông tin về thời gian lưu trữ gói tin trong bộ nhớ (3) của gói tin (2);
- bước 2c: gói tin (2) được tính toán chiều dài bởi bộ tính toán chiều dài gói tin (4) ứng với nguồn dữ liệu (1);
- bước 2d: địa chỉ cơ sở dữ liệu đích của gói tin (2) sẽ được trích xuất bởi bộ tìm kiếm địa chỉ đích (6) ứng với nguồn dữ liệu (1);
- bước 3: các thông tin thu thập được lấy mẫu đều đặn theo thời gian sẽ được gửi đến bộ phân xử chính (18) dưới dạng các tham số đầu vào (7);
- bước 4: các thông số đầu vào (24) sẽ được qua mô hình mạng nơron (19) gồm nhiều lớp: bắt đầu từ lớp tích chập thứ nhất gồm lớp tích chập (25) và hàm kích hoạt (26), qua nhiều lớp tích chập khác đến lớp tích chập cuối cùng gồm lớp tích chập (27) và hàm kích hoạt (28), sau đó là một lớp kết nối đầy đủ (29) để tính được trọng số thực tế (31), lớp tích chập và lớp kết nối đầy đủ sẽ sử dụng các thông số sơ cấp (30) để tính toán ra giá trị sử dụng cho các lớp tiếp theo;
- bước 5: thông số thực tế (31) ứng với gói tin (2) sẽ được đưa đến bộ xử lý ưu tiên (32), nếu trọng số của nó cao hơn tất cả các trọng số của các gói tin khác, gói tin (2) sẽ được ưu tiên gửi đi trước;
- bước 6: sau khi được gửi ra ngoài bộ phân xử chính (18) và trở thành gói tin mới (16), gói tin mới này sẽ được chuyển tới cơ sở dữ liệu đích (17) của nó;
- bước 7a: bộ tính toán tốc độ dữ liệu (13) ứng với cơ sở dữ liệu đích (17) so sánh mức chênh lệch thời gian đôi một giữa gói tin hiện tại (16) và các gói tin trước đó nhận được tại cơ sở dữ liệu đích (17) này và tính toán được tốc độ dữ liệu của cơ sở dữ liệu đích (17);

- bước 7b: bộ đánh giá dữ liệu giữa các cơ sở dữ liệu đích (12) lấy thông tin từ bộ tính toán tốc độ dữ liệu (13) ứng với cơ sở dữ liệu đích (17), cũng như thông tin từ bộ tính toán tốc độ dữ liệu ứng với các cơ sở dữ liệu đích khác (14); từ đó đưa ra thông tin phản hồi về việc chỉnh sửa trọng số ứng với từng cơ sở dữ liệu đích;
- bước 7c: bộ đánh giá dữ liệu giữa các nguồn dữ liệu (10) lấy thông tin từ bộ tính toán tốc độ dữ liệu (5) ứng với nguồn dữ liệu (1), cũng như thông tin từ bộ tính toán tốc độ dữ liệu ứng với các nguồn dữ liệu khác (11); từ đó đưa ra thông tin phản hồi về việc chỉnh sửa trọng số ứng với từng nguồn dữ liệu;
- bước 7d: bộ tính toán chênh lệch thời gian lưu trữ gói tin (8) sẽ tính toán mức độ chênh lệch về thời gian lưu trữ gói tin của gói tin hiện tại (2) so với gói tin trước đó; sau đó so sánh mức chênh lệch này với những nguồn dữ liệu khác (11); từ đó đưa ra thông tin phản hồi về việc chỉnh sửa trọng số ứng với từng nguồn dữ liệu;
- bước 7e: bộ tính toán chênh lệch chiều dài gói tin (9) sẽ tính toán mức độ chênh lệch về chiều dài gói tin còn lại cần xử lý của gói tin hiện tại (2), gói này đã xử lý một phần chiều dài, so với gói tin trước đó; sau đó so sánh mức chênh lệch này với những nguồn dữ liệu khác (11); từ đó đưa ra thông tin phản hồi về việc chỉnh sửa trọng số ứng với từng nguồn dữ liệu;
- bước 8: các thông tin phản hồi thu thập được lấy mẫu đều đặn theo thời gian sẽ được gửi đến bộ phân xử chính (18) dưới dạng các thông số phản hồi (15);
- bước 9: các thông số phản hồi (33) được đưa vào các bộ lan truyền ngược đạo hàm: bắt đầu từ bộ lan truyền ngược của lớp kết nối đầy đủ (34), bộ lan truyền ngược của lớp tích chập cuối cùng (35) theo chiều tính toán trọng số thực tế (31), qua nhiều bộ lan truyền ngược của các lớp tích chập khác, kết thúc ở bộ lan truyền ngược của lớp tích chập đầu tiên (36) theo chiều tính toán trọng số thực tế (31), ở từng bộ lan truyền ngược này, các thông số sơ cấp (30) ứng với từng lớp sẽ được cập nhật lại giá trị mới phù hợp với các thông tin phản hồi (33);

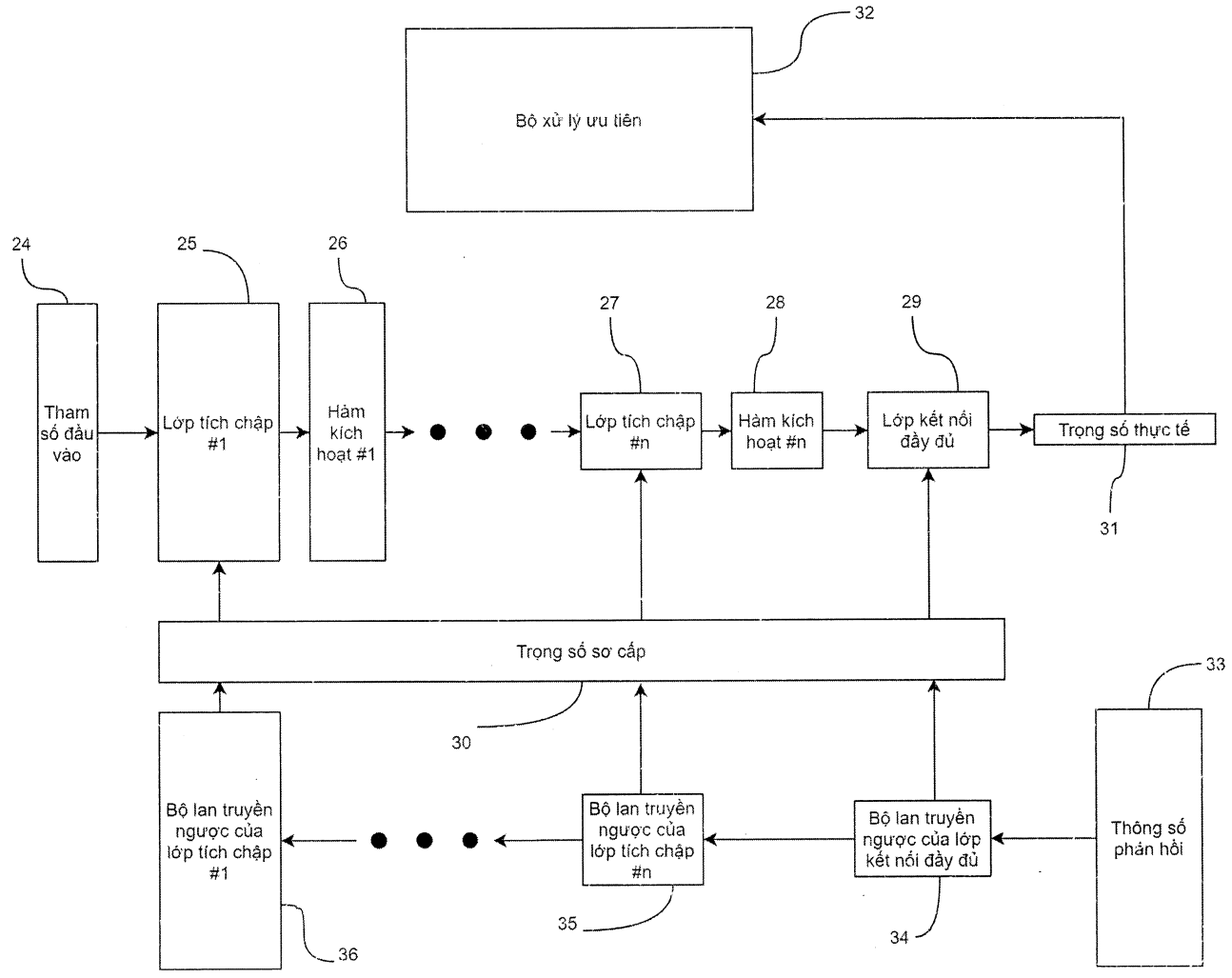
  trong đó:

  các tham số đầu vào và các thông số phản hồi sẽ được lấy mẫu sau mỗi chu kì thời gian nhất định, thời gian lấy mẫu bộ phân xử chính (18), sau mỗi lần lấy mẫu, thông tin nào chưa được cập nhật mới thì nó sẽ được cập nhật trong lần lấy mẫu tiếp theo, thời gian lấy mẫu cần đủ dài để cập nhật thông tin hiệu quả, cũng

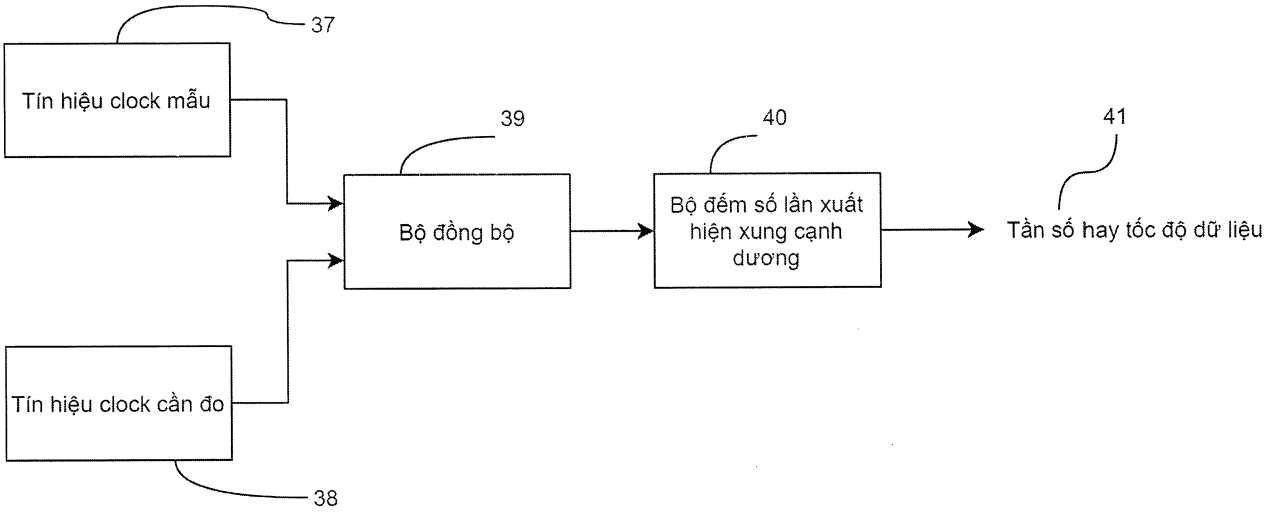
như dành thời gian cho quá trình tính toán trọng số thực tế (22) và quá trình lan truyền ngược.


Thời gian lưu trữ
Bộ tính toán chênh lệch thời gian lưu trữ
Bộ tính toán độ dài gói tin
Bộ tính toán chênh lệch độ dài gói tin
Bộ tính toán tốc độ dữ liệu
Bộ đánh giá các nguồn dữ liệu
Các nguồn dữ liệu khác
Các cơ sở dữ liệu đích khác
Địa chỉ cơ sở dữ liệu đích
Bộ đánh giá các cơ sở dữ liệu đích
Bộ tính toán tốc độ dữ liệu
Tham số đầu vào
Mạng nơ-ron
Thông số phản hồi
Trọng số sơ cấp
Trọng số thực tế
Nguồn dữ liệu
Gói tin
Gói tin
Cơ sở dữ liệu đích
Các gói tin khác
Bộ xử lý ưu tiên
Bộ phân xử chính
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

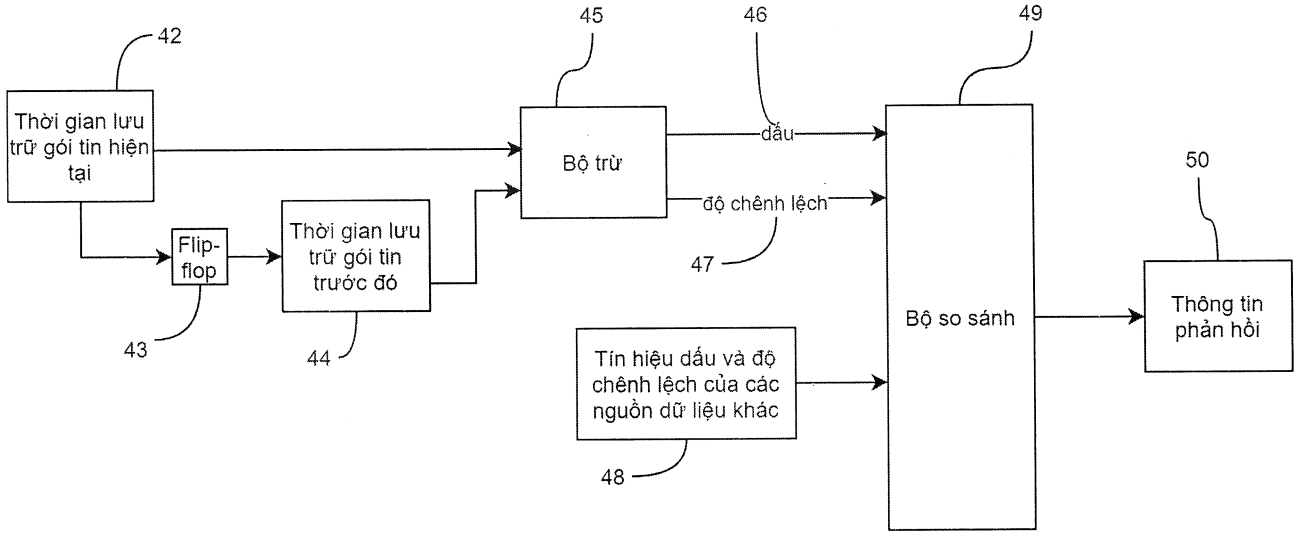
Hình 1


Bộ xử lý ưu tiên
Tham số đầu vào
Lớp tích chập #1
Hàm kích hoạt #1
Lớp tích chập #n
Hàm kích hoạt #n
Lớp kết nối đầy đủ
Trọng số thực tế
Trọng số sơ cấp
Bộ lan truyền ngược của lớp tích chập #1
Bộ lan truyền ngược của lớp tích chập #n
Bộ lan truyền ngược của lớp kết nối đầy đủ
Thông số phản hồi
24
25
26
27
28
29
30
31
32
33
34
35
36

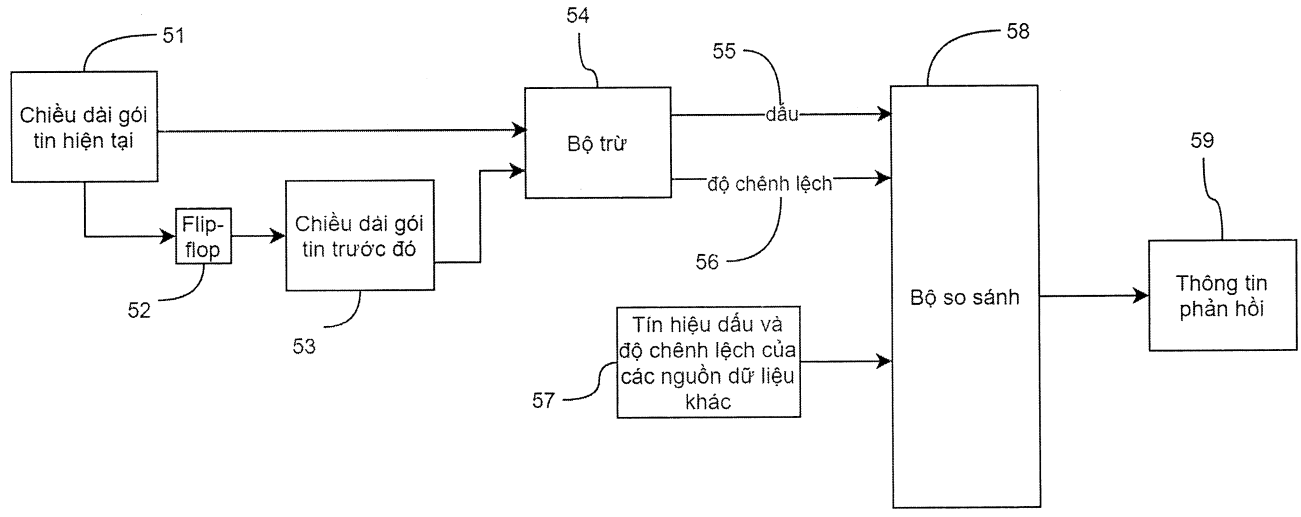
Hình 2



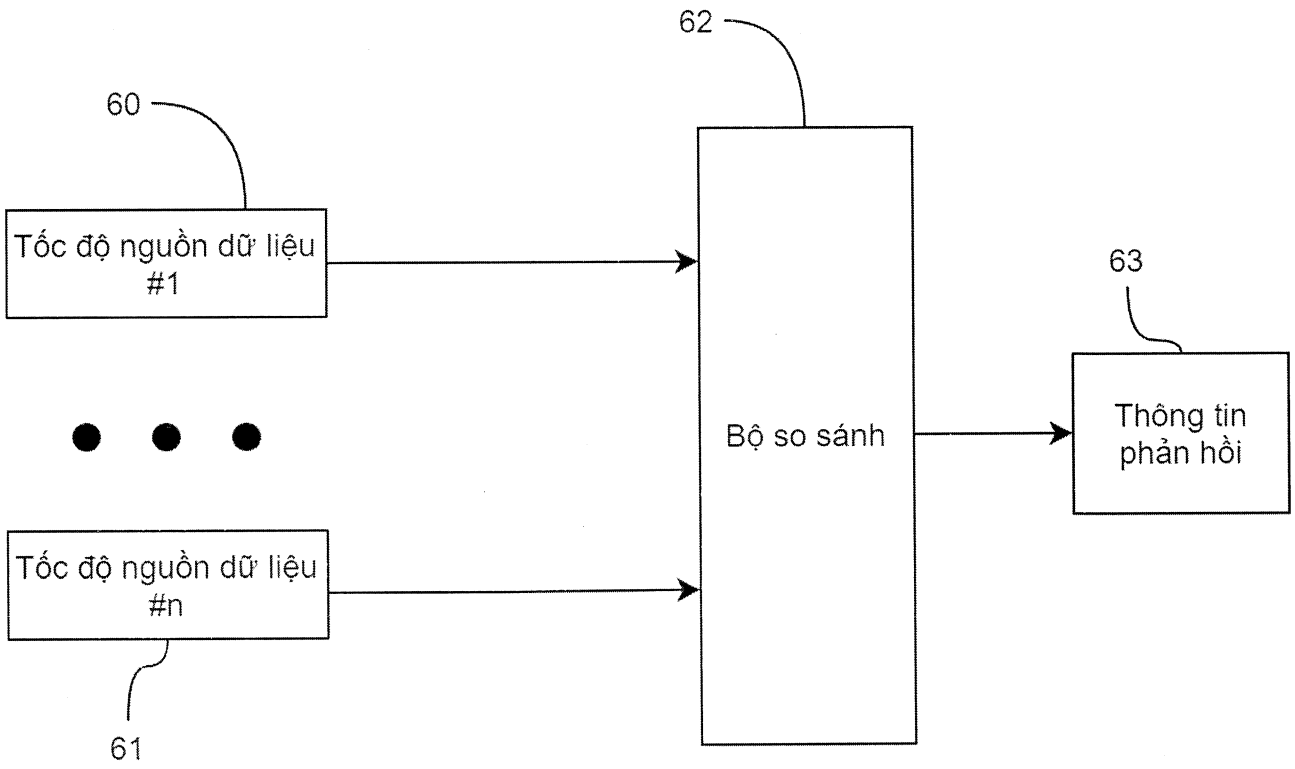
Hình 3



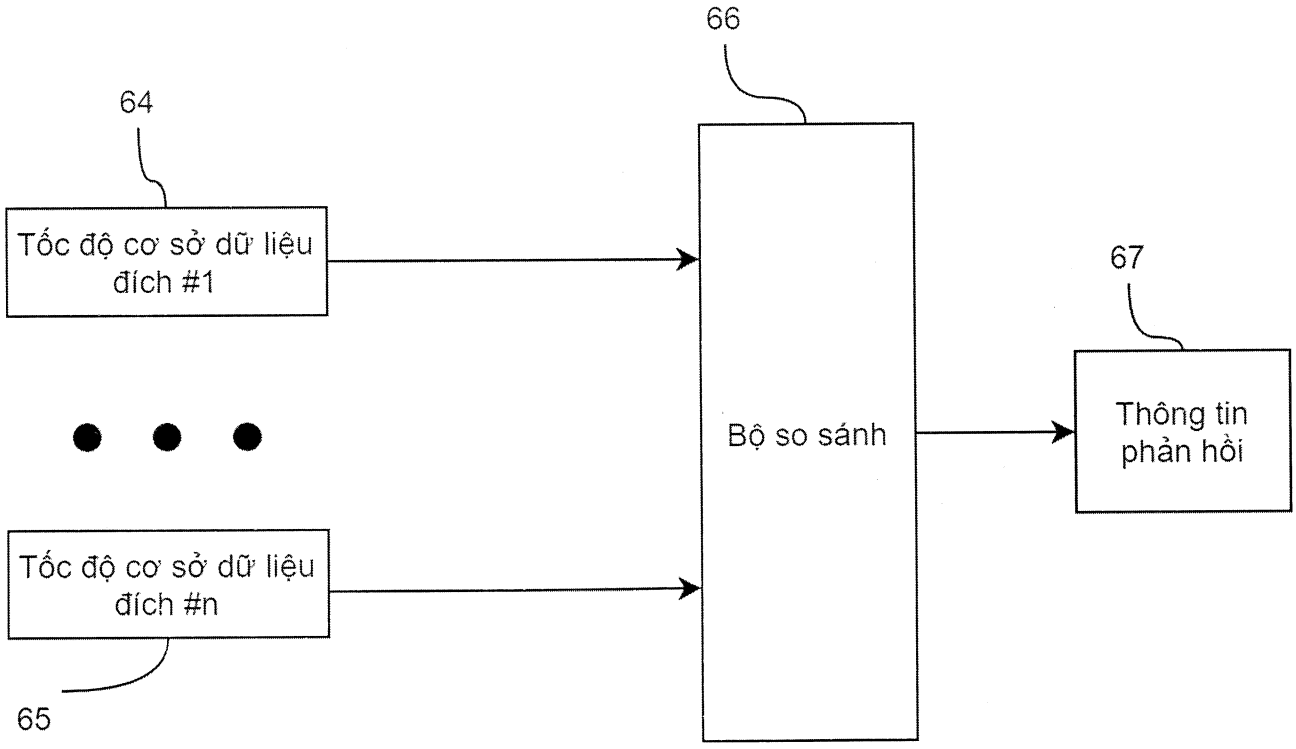
Hình 4



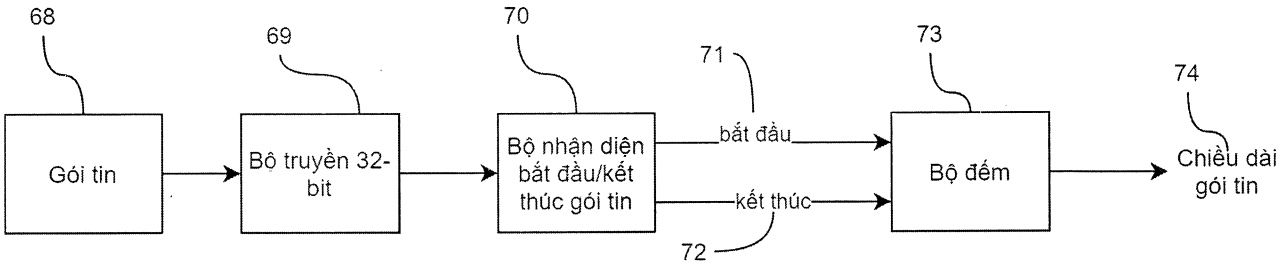
Hình 5


62
60
Tốc độ nguồn dữ liệu #1
63
Bộ so sánh
Thông tin phản hồi
Tốc độ nguồn dữ liệu #n
61

Hình 6


66
64
Tốc độ cơ sở dữ liệu đích #1
67
Bộ so sánh
Thông tin phản hồi
Tốc độ cơ sở dữ liệu đích #n
65

Hình 7


68
Gói tin
69
Bộ truyền 32-bit
70
Bộ nhận diện bắt đầu/kết thúc gói tin
71
bắt đầu
kết thúc
72
73
Bộ đếm
74
Chiều dài gói tin

Hình 8