PTN Tính toán Hiệu năng cao - Khoa KH&KT MT
(HPC Lab - CSE-HCMUT)

CHỦ ĐỀ NGHIÊN CỨU

(TTTN & LUẬN VĂN TỐT NGHIỆP - Cao học & chủ đề mở cho SV khoá sau) 2022

Điều kiện làm việc:

- Sinh viên được cho tham gia các khoá đào tạo ngắn theo chuyên đề để có kiến thức nền liên quan, tiếp cận công nghệ tiên tiến nhất;
- Sinh viên được tạo điều kiện sử dụng hệ thống máy tính tiên tiến tại Trung tâm Tính toán Hiệu năng cao, tài nguyên đặc biệt theo nhu cầu của bài toán;
- Sinh viên được tiếp cận các tập dữ liệu thực tế để có kinh nghiệm trong giải bài toán thực;
- Sinh viên được tạo điều kiện để tham gia các đề tài nghiên cứu cấp quốc gia đang triển khai cũng như có cơ hội trãi nghiệm thực tế;
- Sinh viên làm có kết quả tốt được hỗ trợ công bố kết quả tại các hội nghị, tạp chí khoa học chuyên ngành và trao đổi hợp tác cùng các đơn vị đối tác ở nước ngoài;
- Sinh viên được tham gia các hoạt động học thuật tại PTN; trao đổi cùng các chuyên gia trong và ngoài nước;
- Có kinh phí hỗ trợ mua sắm thiết bị nghiên cứu và các hoạt động nghiên cứu khác;
- Tạo điều kiện cho sinh viên có định hướng nghiên cứu lên Cao học, Tiến sĩ cũng như sinh viên có định hướng làm ở công nghiệp.

Yêu cầu:

- Sinh viên có động lực học tập;
- Không đòi hỏi kiến thức liên quan đến chủ đề vì sinh viên sẽ được đào tạo;
- Có tinh thần làm việc nhóm;
- Trung thực, thẳng thắn, ham thích học hỏi;
- Triển khai công việc theo khả năng và thời gian từng cá nhân nhưng cân đối với tiến độ chung của nhóm;
- Báo cáo đúng hạn;
- Tham gia các buổi seminar để học hỏi thêm và trao đổi quan điểm.

Liên hệ: TS. Nguyễn Quang Hùng, E-mail: nqhung@hcmut.edu.vn

Bài toán 1:

- Tiếng Việt: **Sử dụng** cảm biến ảo thay thế cảm biến thực – Bài toán tối ưu số lượng cảm biết ảo.
- Tiếng Anh:

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2: TS. Nguyễn Quang Hùng
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:
Virtual sensor dùng cho các mục đích sau:

- Virtual sensor dùng thay thế các sensor đắc tiền.
- Sensor không thể triển khai khắp nơi ở diện rộng. Ví dụ ta không thể có sensor quan trắc ở khắp mọi điểm mà chỉ đặt một số nơi và nội suy ra giá trị ở các nơi khác.
- Chúng ta có thể tổng hợp các giá trị đo đạc khác nhau để tạo một ước tính một giá trị đo mới (không thể đo trực tiếp hoặc physical sensor này rất đắc). Ví dụ chúng tác có thể chuyển các đại lượng đo vật lý và hoá học thành tính hiệu điện.
- Chúng ta có thể tạo lại các giá trị đo thực tế trong quá khứ bằng các thuật toán về Machine Learning & thống kê.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu và viết một tài liệu kỹ thuật về virtual sensor.
- Tìm hiểu các phần mềm, công cụ thực hiện virtual sensor; tạo nhiều virtual sensor.
- Nghiên cứu các thuật toán về Machine Learning, thống kê để tổng hợp các giá trị đo đạc khác nhau và tạo ra tính hiệu tổng hợp (lý thuyết chung & ứng dung trong trường hợp cụ thể)
- Nghiên cứu các thuật toán về Machine Learning, thống kê để sinh dữ liệu từ dữ liệu quá khứ; ứng dụng trong trường hợp dữ liệu quan trắc trong nông nghiệp
- Xây dựng kịch bản xây dựng tạo hệ thống quan trắc ảo dùng virtual sensor trong nông nghiệp.
- Triển khai thử nghiệm giải pháp (có điều kiện triển khai thực tế & viết bài báo khoa học);
- Thử nghiệm, đánh giá và cải tiến.

Tài liệu tham khảo:

[1] https://link.springer.com/content/pdf/10.1007/s12599-021-00689-w.pdf.

Bài toán 2:

- Tiếng Việt: Khung phần mềm thu thập dữ liệu IoT dùng lõi ThingsBoard
- Tiếng Anh: An IoT data collection platform using ThingsBoard core

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2: TS. Nguyễn Quang Hùng
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:
Xu hướng hướng dữ liệu (data-driven) là giải pháp tất yếu, tiên tiến để giải quyết nhiều bài toán. Một khâu quan trọng là thu thập dữ liệu. Có nhiều nguồn dữ liệu để thu thập như: hệ thống cảm biến (sensor), mạng xã hội, camera, mobile phone, v.v.. Trong đó dữ liệu từ sensor ngày càng nhiều cùng với sự phát triển của IoT & 5G. Bài toán chính là phát triển một khung phần mềm thu thập dữ liệu IoT.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu các giao thức MQTT, CoAP, HTTP, 4G/5G;
- Tìm hiểu thư viện MQTT [2], ThingsBoard [1], Kafka [3] và viết report về các khung phần thu thập dữ liệu ThingsBoard & thư viện MQTT;
- Tìm hiểu chi tiết các API và kiến trúc của ThingsBoard [3] (môi trường thử nghiệp được triển khai trên hệ thống máy SuperNode-XP tại HPC Lab);
- Tìm hiểu và đề xuất giải pháp định danh, quản lý vòng đời của từng sensor kết nối;
- Tìm hiểu và phát triển các giải thuật tiền xử lý dữ liệu (có hỗ trợ trong ThingsBoard): filtering, sampling, integration;
- Phát triển một khung phần mềm thu thập dữ liệu hỗ trợ MQTT, CoAP, HTTP, 4G/5G dự trên lõi của ThingsBoard;
- Triển khai thử nghiệm giải pháp: có điều kiện triển khai thực tế thu thập dữ liệu nông nghiệp, giao thông;
- Thử nghiệm, đánh giá và cải tiến;
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] ThingsBoard: https://thingsboard.io.
[2] MQTT: https://mqtt.org.
[3] Kafka: https://kafka.apache.org.

Bài toán 3:

- Tiếng Việt: Khung phần mềm thu thập và tích hợp dữ liệu trung tâm
- Tiếng Anh: A central data collection & integration platform

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2: TS. Nguyễn Quang Hùng
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:
Hướng dữ liệu (data-driven) là giải pháp tất yếu, tiên tiến để giải quyết nhiều bài toán. Một khâu quan trọng là thu thập dữ liệu. Có nhiều nguồn dữ liệu để thu thập như: hệ thống cảm biến (sensor), mạng xã hội, camera, mobile phone, v.v.. Việc kết nối tập trung tất cả các sensor hay các nguồn tạo dữ liệu khác nhau về một trung tâm sẽ gây tắc nghẽn đường truyền cũng như quá tải tại nút trung tâm; một kiến trúc thu thập dữ liệu phân tán theo mô hình tính toán biên (edge computing) hay tính toán sương mù (fog computing) là giải pháp hợp lý. Việc thu thập dữ liệu tại nút biên xem như đã được giải quyết (ở Bài toán 2), Bài toán 3 này tập trung giải quyết về kiến trúc phân tán và thu thập dữ liệu cho nhiều ứng dụng khác cùng chia sẻ hạ tầng thu thập dữ liệu.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu các giao thức MQTT, CoAP, HTTP, 4G/5G, và các giao thức (protocol hỗ trợ trong Kafka) (tham khảo thêm [4]).
- Tìm hiểu thư viện MQTT [2], ThingsBoard [1], Kafka [3] và viết report về các khung phần thu thập dữ liệu Kafka;
- Tìm hiểu chi tiết các API và kiến trúc của Kafka [3][4][5] (môi trường thử nghiệp được triển khai trên hệ thống máy SuperNode-XP tại HPC Lab);
- Thiết kế kiến trúc thu thập dữ liệu phân tán cho nhiều ứng dụng khác nhau cùng sử dụng (chia sẻ) một hạ tầng thu thập dữ liệu, tham khảo [6] [7];
- Giải pháp xác thực, kết nối, quản lý các nút kết nối trong kiến trúc thu thập dữ liệu phân tán (có ứng dụng yêu cầu dữ liệu riêng biệt nhưng có một số ứng dụng có thể chia sẻ dữ liệu với nhau để làm giàu dữ liệu có được);
- Giải pháp thu nhận, phân chia, tiền xử lý dữ liệu tại các nút trung tâm sử dụng lõi Kafka;
- Phát triển một khung phần mềm thu thập dữ liệu phân tán hỗ trợ nhiều ứng dụng trên lõi của Kafka;
- Triển khai thử nghiệm giải pháp: có điều kiện triển khai thực tế thu thập dữ liệu nông nghiệp, giao thông, y tế và giáo dục;
- Thử nghiệm, đánh giá và cải tiến;
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] ThingsBoard: https://thingsboard.io.
[2] MQTT: https://mqtt.org.
[3] Kafka: https://kafka.apache.org.

[4] https://www.wowza.com/blog/streaming-protocols.
[5] Kafka Event Streaming Platform https://www.confluent.io/blog/event-streaming-platform-1/
[6] Big Data Architecture https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/
[7] Node.js framework server-side applications. https://nestjs.com/

Bài toán 4:

- Tiếng Việt: Khung phần mềm phân tích dữ liệu dùng lõi Spark
- Tiếng Anh: A data analytics platform using Spark

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2: TS. Nguyễn Quang Hùng
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:
Xu hướng hướng dữ liệu (data-driven) là giải pháp tất yếu, tiên tiến để giải quyết nhiều bài toán. Có nhiều nguồn dữ liệu để thu thập như: hệ thống cảm biến (sensor), mạng xã hội, camera, mobile phone, v.v.. và dữ liệu thu thập là dữ liệu lớn ở dạng có cấu trúc và phi cấu trúc. Dữ liệu được thu thập theo dạng dòng (stream). Đòi hỏi một khung phần mềm có khả năng phân tích dữ liệu với nhiều định dạng khác nhau cũng như dữ liệu dạng dòng (data streaming). Spark là một công cụ phân tích dữ liệu mạnh mẽ được sử dụng nhiều trong công nghiệp Google, Facebook, Amazone, v.v. với các module về: SQL, Streaming, Machine Learning, GraphX; hỗ trợ ngôn ngữ Scala, Java & Python cũng nhiều có nhiều dự án của đối tác khác (Third-party projects). Bài toán chính là phát triển một khung phần mềm có khả năng phân tích dữ liệu dựa trên Spark.

Một phần kết quả đã được thực hiện tốt trong LVTN đại học của Nguyễn Văn Hoài Linh. Công việc trong nghiên cứu này là tiếp tục phát triển những tính năng mới nên có nhiều thuận lợi.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu Spark [1] và viết report kỹ thuật;
- Tìm hiểu Kafka [2] để nhận dữ liệu streaming từ Kafka về Saprk;
- Đọc tài liệu, triển khai lại giải pháp tích hợp một hàm mới về Fuzzy logic vào Spark [3] (môi trường thử nghiệp được triển khai trên hệ thống máy SuperNode-XP tại HPC Lab);
- Phát triển phát triển một khung phần mềm có khả năng phân tích dữ liệu dựa trên Spark;
- Triển khai thử nghiệm giải pháp: có điều kiện triển khai thực tế thu thập dữ liệu nông nghiệp, giao thông, y tế, giáo dục;
- Thử nghiệm, đánh giá và cải tiến;
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] Spark: https://spark.apache.org.
[2] Kafka: https://kafka.apache.org.
[3] Luận văn tốt nghiệp của Nguyễn Văn Hoài Linh năm 2021.

Bài toán 5:

- Tiếng Việt: Kiến trúc Lakehouse cho dữ liệu lớn về Giao thông
- Tiếng Anh: Big data architecture

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2: TS. Nguyễn Quang Hùng
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:
Xu hướng hướng dữ liệu (data-driven) là giải pháp tất yếu, tiên tiến để giải quyết nhiều bài toán. Có nhiều nguồn dữ liệu để thu thập như: hệ thống cảm biến (sensor), mạng xã hội, camera, mobile phone, v.v.. và dữ liệu thu thập là dữ liệu lớn ở dạng có cấu trúc và phi cấu trúc. Dữ liệu được thu thập theo dạng dòng (stream). Kiến trúc tổng thể xử lý dữ liệu lớn khác với mô hình truyền thống sử dụng Data Warehouse với kỹ thuật trích xuất, biến đổi và tải ETL. Một số mô hình mới được đề xuất như Lambda [3], Kappa [6] & Delta [4-5]. Bài toán chính là phát triển một giải pháp triển khai Delta (tham khảo Lambda, Kappa) dựa trên công cụ mã mở như Kafka [2], Spark [1] & Delta Lake [5].

Sinh viên tham gia có cơ hội làm việc trên dữ liệu thực về Giao thông của Thành phố và các tỉnh thành khác.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu về kiến trúc Lambda, Kappa và Delta; viết report về các kiến trúc này;
- Tìm hiểu Spark [1], Kafka [2], Lakehouse [5];
- Đọc tài liệu, triển khai lại giải pháp Delta (môi trường thử nghiệp được triển khai trên hệ thống máy SuperNode-XP tại HPC Lab);
- Triển khai thử nghiệm giải pháp: có điều kiện triển khai thực tế thu thập dữ liệu nông nghiệp, giao thông, y tế, giáo dục;
- Thử nghiệm, đánh giá và cải tiến;
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] Spark: https://spark.apache.org.
[2] Kafka: https://kafka.apache.org.
[3] Lambda: "Lambda Architecture - Realtime Data Processing", http://dx.doi.org/10.13140/RG.2.2.19091.84004.
[4] Lakehouse with Delta architecture: http://cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf (https://cs.stanford.edu/~matei/papers/2021/cidr_lakehouse.pdf).
[5] Opensoure Lakehouse: https://delta.io.
[6] Kappa & Lambda: (1) https://towardsdatascience.com/a-brief-introduction-to-two-data-processing-architectures-lambda-and-kappa-for-big-data-4f35c28005bb ; (2) https://en.paradigmadigital.com/dev/from-lambda-to-kappa-evolution-of-big-data-architectures/.

Bài toán 6:

- Tiếng Việt: Phần mềm nền giám sát hệ thống & ứng dụng
- Tiếng Anh: Smart monitoring tools

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2: TS. Nguyễn Quang Hùng
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:
Những hệ thống tính toán hiệu năng cao (HPC - High Performance Computing) hay thu thập và phân tích dữ liệu (lớn) (Data collection & analytics system) trên một và nhiều máy tính đều có nhiều phần tử chức năng chạy trên máy tính thực, máy tính ảo (VMs) và Docker. Một công cụ giám sát cho những hệ thống này là không thể thiếu. Một thế hệ công cụ mới dựa trên thiết kế theo dạng microservice sử dụng thư viện MQTT [1] hay Kafka [2] là đang được quan tâm. DCDB [3] là một công cụ giám sát như vậy được phát triển và sử dụng tại Trung tâm Siêu máy tính Leibniz - Đức (https://www.lrz.de/english/). Virtual sensor là các cảm biến phần mềm/phần cứng để cung cấp thông tin trên từng nút tính toán và các thông tin trên các nút tính này được phân cấp để truyền về nút trung tâm [3]. Wintermute [4] là một công tích hợp với DCDB sử dụng thông tin giám sát của DCDB để tiến hành phân tích và ra quyết định điều khiển theo các luật (giải thuật) được cài đặt linh động. Bài toán này hướng đến sử dụng DCDB & Wintermute để (1) giám sát hoạt động của phần mềm nền tảng (platform) khi hoạt động và (2) phát triển các giải thuật phân tích và điều khiển thông minh.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu công cụ giám sát như Prometheus [6], Zabbix [7], đặc biệt là DCDB [3] và viết report;
- Tìm hiểu Wintermute [4] và viết report;
- Triển khai DCDB & Wintermute [5] trên hệ thống5 máy tính SuperNode-XP tại HPC Lab;
- Phát triển một phần mềm nền (monitoring platform) dựa trên DCDB & Wintermute cho ĐHBK.
- Phát triển các virtual sensor để giám sát dịch vụ cho Spark [1], Kafka [2] và phần mềm nền khác;
- Phát triển một phần mềm nền (monitoring platform) dựa trên DCDB & Wintermute cho ĐHBK, hệ thống phần mềm nền khác như Smart village platform;
- Triển khai thử nghiệm giải pháp: có điều kiện triển khai thực tế trên SuperNode-XP, ĐHBK & Smart village platform;
- Thử nghiệm, đánh giá và cải tiến;
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] MQTT: https://mqtt.org.

[2] Kafka: https://kafka.apache.org.
[3] DCDB: From Facility to Application Sensor Data: Modular, Continuous and Holistic Monitoring with DCDB: https://arxiv.org/abs/1906.07509.
[4] DCDB Wintermute: Enabling Online and Holistic Operational Data Analytics on HPC Systems: https://arxiv.org/pdf/1910.06156.pdf.
[5] DCDB source: https://gitlab.lrz.de/dcdb/dcdb.
[6] Prometheus: https://prometheus.io.
[7] Zabbix: https://www.zabbix.com.

Bài toán 7:

- Tiếng Việt: Giải pháp xử lý dữ liệu lớn CDNs.
- Tiếng Anh: Big data analytics on CDNs.

CBHD1: PGS. TS. Thoại Nam

Email1: namthoai@hcmut.edu.vn

CBHD2:

Email1: nqhung@hcmut.edu.vn

Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:
Sự phát triển nhanh của Internet, thiết bị mobile phone, công nghệ truyền dẫn 4G/5G đẩy mạnh sự phát triển nội dung số nhất là lĩnh vực đa phương tiện ứng dụng trong giải trí, học tập, v.v.. Khi số lượng người sử dụng đầu cuối lớn thì giải pháp phân phối nội dung trên mạng (Content Delivery Network - CDN) bắt buộc phải ứng dụng tại các nhà cung cấp dịch vụ Internet (Internet Service Provider - ISP) nếu tham gia thị trường phân phối nội dung số. Bảo chất lượng dịch vụ và tiết kiệm tài nguyên cũng như băng thông mạng bên trong của các ISP thì có hai bài toán chính cần giải quyết: (1) Giải pháp và giải thuật Caching, và (2) Chiến lược sử dụng tài nguyên hợp lý trên nền điện toán đám mây phù hợp với tải sử dụng (thay đổi theo giờ).

Bài toán 1 có hai giải thuật LRU và Color-based Caching. Bài toán 2 thì có hai luận văn đại học và cao học (ThS. La Hoàng Lộc) phát triển một công cụ giả lập (emulator) trên Mininet, giải pháp tìm lời giải tối ưu sử dụng Gaussian Process Regression & Bayesian Optimization (đã có một số kết quả công bố ở các hội nghị uy tín IEEE/ACM/Springer-Verlag). Nghiên cứu này tiếp bước các kết quả đã có nên rất thuận lợi.

Đối với bài toán sử dụng tài nguyên cho phù hợp một cách tiếp cận phổ biến nhất là quy về bài toán tối ưu hóa. Tuy nhiên, tùy thuộc vào từng bài toán cụ thể mà bài toán này sẽ được mô hình hóa như bài toán đa hay đơn mục tiêu. Bài toán cụ thể được quan tâm trong luận văn này là bài toán replacement caching với 2 mục tiêu chính là tối đa hóa chất lượng dịch vụ và tối thiểu hóa chi phí đầu tư (số lượng replica servers trong mạng). Để giải quyết bài toán này, hướng tiếp cận truyền thống thường sẽ sử dụng toán để mô hình hóa bài toán và tính toán hàm mục tiêu. Tuy nhiên hướng tiếp cận này thường sẽ đặt ra một số giả định phi thực tế nhằm đơn giản hóa bài toán. Một hướng tiếp cận khác là sử dụng một công cụ emulator để tính toán hàm mục tiêu này tận dụng khả năng giả lập gần giống môi trường thực của công cụ. Luận văn này có 2 mục tiêu chính: so sánh các giải thuật mô hình hoán toán học [1] với hướng tiếp cận sử dụng emulator kết hợp với giải thuật tối ưu hóa Bayesian[3], và đề xuất giải thuật tối ưu hóa khác phù hợp hơn.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu CDNs, các giải thuật caching thông dụng.
- Tìm hiểu về các mô hình toán truyền thống nhằm tối ưu hóa bài toán replica server placement. [1]
- Tìm hiểu về công cụ giả lập emulator đã được lab phát triển. [2]
- Tìm hiểu về hướng tiếp cận sử dụng giải thuật Bayesian với công cụ emulator để tối ưu hóa bài toán replica server placement. [3]

- Thử nghiệm, đánh giá và cải tiến;
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] J. Sahoo, M. A. Salahuddin, R. Glitho, H. Elbiaze and W. Ajib, "A Survey on Replica Server Placement Algorithms for Content Delivery Networks," in *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1002-1026, Secondquarter 2017.

[2] H. -L. La, A. -T. N. Tran, M. Yoshimi, T. Nakajima and N. Thoai. “CDNET: A Content Delivery Network Emulator." In 2021 International Symposium on Networks, Computers and Communications (ISNCC), 2021.

[3] H. -L. La, H-Thanh Hoang Le, and Nam Thoai. "A Multi-Objective Approach for Optimizing Content Delivery Network System Configuration,"

Bài toán 8:

- Tiếng Việt: Giải pháp lập lịch hiệu năng cao ứng dụng AI.
- Tiếng Anh: High performance Scheduling using AI.

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2:
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:

Hệ thống máy tính tính toán hiệu năng ngày càng được sử dụng rộng rãi do xu thế ứng dụng tính toán và phân tích dữ liệu trong nhiều lĩnh vực khác. Các hệ thống này đắc tiền nên việc sử dụng chúng hiệu quả là một ưu tiên hàng đầu.

Bài toán lập lịch thực thi ứng dụng có vai trò quan trọng quyết định hiệu suất sử dụng tài nguyên của hệ thống máy tính hiệu năng cao. Tuy nhiên, với cách vận hành hiện tại thì bài toán này phụ thuộc không chỉ giải thuật lập lịch mà còn phụ thuộc vào đặc tính của ứng dụng dựa trên các thông số của người sử dụng nhập vào hệ thống như thời gian chạy dự kiến của ứng dụng. Không may là đa số trường hợp thì người sử dụng lại nhập thông số này không chính xác vì nhiều nguyên do. Bài toán này nghiên cứu và phát triển giải pháp giải quyết vấn đề này trên hai nội dung chính: (1) Thuật toán dự đoán thời gian thực thi ứng dụng của người sử dụng khi đăng ký công việc dựa trên thuật toán học máy kNN, và (2) Giải thuật lập lịch ứng dụng kỹ thuật học sâu tăng cường (Deep Reinforcement Learning). Các nội dung này được triển khai đánh giá trên tập dữ liệu thực từ bốn trung tâm siêu máy tính trên thế giới cộng với tập dữ liệu thực tế có trên hệ thống SuperNode-XP tại Trường Đại học Bách Khoa - ĐHQG-HCM.

Bài toán trên đã được giải quyết một phần ở một luận văn cao học (ThS. Hoàng Lê Hải Thanh) và đã có một số kết quả công bố ở các hội nghị uy tín IEEE/ACM/Springer-Verlag. Nghiên cứu này tiếp bước các kết quả đã có nên rất thuận lợi.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu về các hệ thống HPC trên thế giới và tại Việt Nam như SuperNode-XP
- Tìm hiểu về các tập dữ liệu lịch sử hoạt động của các hệ thống HPC, các định dạng workload phổ biến như SWF [1]
- Phân tích, đánh giá hiện trạng sử dụng tài nguyên và hành vi của người dùng trên các hệ thống HPC từ các tập dữ liệu [2][3]
- Nghiên cứu các kỹ thuật học máy như Linear Regression, kNN, SVM, Random Forest hay Deep Learning [4]
- Thiết kế các mô hình học máy nhằm mục tiêu dự đoán các thông số sử dụng tài nguyên của các công việc, tham khảo các phương pháp đã được đề xuất [5]
- Tìm hiểu lý thuyết mô hình học tăng cường, các phương pháp huấn luyện học tăng cường như A2C, PPO [6]
- Nghiên cứu xây dựng cách biểu diễn trạng thái các hệ thống HPC cho mô hình DRL, tham khảo các mô hình đã được đề xuất như DeepRM [7]

- Thiết kế mô hình học sâu tăng cường giúp ra quyết định định thời công việc trên các hệ thống HPC
- Triển khai huấn luyện và kiểm thử mô hình bằng các ML framework như PyTorch [8], TensorFlow [9] hoặc Theano trên hệ thống SuperNode-XP hoặc Google Colab
- Thử nghiệm, so sánh các giải pháp
- Đánh giá và cải tiến;
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] Parallel Workloads Archive: https://www.cs.huji.ac.il/labs/parallel/workload/

[2] Workload Characterization and Modeling Book: https://www.cs.huji.ac.il/~feit/wlmod/

[3] Experience with using the Parallel Workloads Archive https://www.cs.huji.ac.il/~feit/papers/PWA14JPDC.pdf

[4] Ebook "Machine Learning cơ bản", Vũ Hữu Tiệp: https://github.com/tiepvupsu/ebookMLCB/blob/master/book_ML_color.pdf

[5] Potential of Applying kNN with Soft Walltime to Improve Scheduling Performance: https://www.researchgate.net/publication/353451324_Potential_of_Applying_kNN_with_Soft_Walltime_to_Improve_Scheduling_Performance

[6] Sutton & Barto Book: Reinforcement Learning: An Introduction: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

[7] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, “Resource Management with Deep Reinforcement Learning,”: https://people.csail.mit.edu/alizadeh/papers/deeprm-hotnets16.pdf

[8] PyTorch Reinforcement Learning (DQN) Tutorial: https://pytorch.org/tutorials/intermediate/reinforcement_q_learning.html

[9] TensorFlow Introduction to RL and Deep Q Networks: https://www.tensorflow.org/agents/tutorials/0_intro_rl

Bài toán 9:

- Tiếng Việt: Giải pháp tính toán phân tán cho DL (Deep Learning) và Graph Neuron Networks (GNNs).
- Tiếng Anh: Solutions for Distributed Deep Learning (DL) and Graph Neuron Networks (GNNs).

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2: TS. Nguyễn Quang Hùng
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:

Deep Learning (DL) và Graph Neuron Networks (GNNs) được sử dụng ngày càng nhiều nhưng thách thức là độ phức tạp tính toán và bộ nhớ cần tăng dần khi dữ liệu càng lớn; vì vậy giải pháp tính toán phân tán trong huấn luyện (training) của DL & GNNs là cần thiết.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Nghiên cứu đánh giá mô hình song song cho tính toán phân tán trong huấn luyện mô hình DL & GNNs;
- Nghiêu cứu đánh giá kiến trúc máy chủ tham số trong huấn luyện mô hình học sâu;
- Nghiên cứu phương pháp huấn luyện dữ liệu phân tán federated Learning (Collaborative Learning);
- Nghiên cứu giải pháp tính toán phân tán của TensorFlow, PyTorch, Hovorod và triển khai thực tế;
- Phát triển giải pháp tính toán phân tán trên nhiều cụm máy;
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] Distributed Training of Deep Learning Models: A Taxonomic Perspective: *https://arxiv.org/abs/2007.03970*.
[2] Awesome Distributed Deep Learning: *https://github.com/bharathgs/Awesome-Distributed-Deep-Learning*.
[3] DDL in TensorFlow: *https://www.tensorflow.org/guide/distributed_training*.
[4] DDL in PyTorch: *https://pytorch.org/tutorials/beginner/dist_overview.html*.
[5] Horovod AI: *https://horovod.ai/*.
[6] DDL in Horovod: *https://github.com/horovod/horovod*.
[7] Computing Graph Neural Networks: A Survey from Algorithms to Accelerators: *https://arxiv.org/abs/2010.00130*.
[8] TensorFlow GNN: *https://github.com/tensorflow/gnn*.

Bài toán 10:

- Tiếng Việt: Giải pháp tính toán học máy biên trong dữ liệu dòng.
- Tiếng Anh: Edge machine learning for Data streaming.

CBHD1: PGS. TS. Thoại Nam
Email1: namthoai@hcmut.edu.vn
CBHD2: TS. Nguyễn Quang Hùng
Email1: nqhung@hcmut.edu.vn
Họ tên – mã số sinh viên thực hiện (nếu đã có): (5 SV)

Mô tả:

Hướng dữ liệu (data-driven) là giải pháp tất yếu, tiên tiến để giải quyết nhiều bài toán. Một khâu quan trọng là thu thập dữ liệu. Có nhiều nguồn dữ liệu để thu thập như: hệ thống cảm biến (sensor), mạng xã hội, camera, mobile phone, v.v.. Việc kết nối tập trung tất cả các sensor hay các nguồn tạo dữ liệu khác nhau về một trung tâm sẽ gây tắc nghẽn đường truyền cũng như quá tải tại nút trung tâm; một kiến trúc thu thập dữ liệu phân tán theo mô hình tính toán biên (edge computing) hay tính toán sương mù (fog computing) là giải pháp hợp lý. Hướng phân tích dữ liệu ứng dụng Machine learning/Deep learning là một xu thế hiện tại.

Bài toán chính là phát triển một giải pháp Edge machine learning in Data streaming.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu về tính toán trong Data streaming [1];
- Tìm hiểu các ý tưởng giải thuật distributed machine learning có thể ứng dụng trong edge machine learning [2] trong data streaming;
- Tìm hiểu Kafka [3] thư viện truyền nhận phổ dụng trong data streaming;
- Tìm hiểu Esper [4] là thư viện mã nguồn mở hỗ trợ tính toán trong data streaming;
- Tìm hiểu giải pháp tích hợp Esper và Kafka [5];
- Đề xuất một giải thuật edge machine learning (dựa trên [2]);
- Viết báo cáo;
- Phát triển giải thuật edge machine learning trên khung Kafka+Esper (SV có thể chọn lựa môi trường khác để triển khai);
- Viết bài báo khoa học.

Tài liệu tham khảo:

[1] A Comprehensive Survey on Parallelization and Elasticity in Stream Processing : *https://www.semanticscholar.org/paper/A-Comprehensive-Survey-on-Parallelization-and-in-R%C3%B6ger-Mayer/274afbef7369c7a4d91fbfa17be56bbc15cb25ac.*

[2] Basic Scalable Streaming Algorithms: *https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5ae7bc030&appId=PPGMS.*

[3] Edge Machine Learning for AI-Enabled IoT Devices: A Review: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7273223/pdf/sensors-20-02533.pdf*.

[4] Machine Learning at the Network Edge: *A Survey: https://arxiv.org/abs/1908.00080*.

[5] Machine learning for streaming data: state of the art, challenges, and opportunities: *https://www.researchgate.net/publication/337581742_Machine_learning_for_streaming_data_state_of_the_art_challenges_and_opportunities*.

[6] Kafka: *https://kafka.apache.org.*
[7] Esper 2019. Esper. (jan 2019). Retrieved 2019-01-15 from *http://www.espertech.com/*.
[8] Esper & Kafka: *https://github.com/lwluc/Esper-Kafka-Example*.

Các nhóm chọn 1 đề tài, ghi họ tên & MSSV các thành viên trong nhóm lên comment bên dưới. Nhóm không cần có KSTN.

Tên đề tài MR202: Giải thuật phân tích dữ liệu trên Kafka và Spark Streaming cho dữ liệu IoTs

CBHD1: TS. Nguyễn Quang Hùng CBHD2:

Email1: nqhung@hcmut.edu.vn Email2: Email3:

Số lượng sinh viên thực hiện: 5 sinh viên

Mô tả đề tài:

✔ Dữ liệu IoTs rất lớn trong thời đại CN 4.0, GD 4.0. Sinh viên tìm hiểu kiến trúc IoTs, và Apache Kafka và Spark Streaming được đề xuất kết nối với IoTs framework để lưu trữ và phân tích dữ liệu. Sinh viên tìm hiểu và đề xuất giải thuật xử lý trên data streaming từ các IoTs device. Ứng dụng với dữ liệu Làng Thông Minh của PGS. TS. Thoại Nam.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

✔ Tìm hiểu về IoT framewokr như Thingsboard.

✔ Tìm hiểu về machine learning, Big Data, Kafka, Spark.

✔ Đề xuất mô hình và triển khai thử nghiệm Thingsboard framework với Kafka và Spark streaming.

✔ Đề xuất giải thuật xử lý data streaming trên hệ thống.

Tài liệu tham khảo:

✔ https://thingsboard.io/

- ✔ https://kafka.apache.org/
- ✔ http://spark.apache.org/docs/latest/index.html

Tên đề tài MR203: AI-driven resource allocation algorithm for application in cloud and fog computing environment

CBHD1: TS. Nguyễn Quang Hùng CBHD2: PGS. TS. Thoại Nam

Email1: nqhung@hcmut.edu.vn Email2:

Số lượng sinh viên thực hiện: 5 sinh viên

Mô tả đề tài:

✔ Đám mây và fog cung cấp tài nguyên tính toán rất lớn và kỹ thuật ảo hóa hỗ trợ nhiều nhóm khác nhau triển khai các ứng dụng tính toán trên đó. Nhiệm vụ cấp phát tài nguyên hướng AI (sử dụng các kỹ thuật AI) cho các ứng dụng tính toán trên hệ thống đám mây và sương mù hiệu quả về hiệu năng, chi phí.

Cloud and fog provide huge computing resources and virtualization technology supports many different groups deploying compute applications on it. AI-driven resource allocation tasks (using AI techniques) for performance and cost-effective cloud and fog computing applications.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- ✔ Tìm hiểu công nghệ điện toán đám mây, Fog computing.
- ✔ Tìm hiểu về docker, VM, OpenStack.
- ✔ Đề xuất giải thuật cấp phát tài nguyên cho các ứng dụng trên hệ thống cho AI và Big Data.
- ✔ Viết thuật toán mô phỏng.

Tài liệu tham khảo:

- ✔ https://openai.com/
- ✔ Slides và tài liệu môn Tính toán song song.

Tên đề tài MR204: Xây dựng hệ thống phân tích đa luồng IoT thông minh

CBHD1: TS. Nguyễn Quang Hùng CBHD2:

Email1: nqhung@hcmut.edu.vn Email2:

Số lượng sinh viên thực hiện: 5 sinh viên

Mô tả đề tài:

✔ Đề tài tìm hiểu và đề xuất kiến trúc hệ thống xử lý đa luồng dữ liệu IoT thông minh (có cài đặt giải thuật AI/ML để cảnh báo và phân tích tình trạng của bệnh nhân dựa trên dữ liệu thu thập được từ các IoT devices) cho phép gửi nhận dữ liệu từ các thiết bị IoT ở xa qua mạng Wireless/4G/5G.

✔ Dữ liệu từ các IoT devices/IoT gateway gửi về Thingsboard, yêu cầu dữ liệu truyền nhận đạt độ tin cậy cao.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

✔ Tìm hiểu kiến trúc Edge computing cho IoTs devices.

✔ Tìm hiểu Thingsboard và các mô hình kết nối IoT devices/Gateways và Thingsboards.

✔ Xây dựng mô hình mẫu để kiểm nghiệm và đánh giá sơ bộ.

✔ Giai đoạn LVTN: Hoàn thiện mô hình hệ thống IoTs thông minh dựa trên Edge computing đã đề xuất, triển khai với một đơn vị y tế.

Tài liệu tham khảo: Liên hệ GVHD

Tên đề tài (MR205): Thu thập và phân tích hoạt động học tập của người học trên các LMS

CBHD1: Nguyễn Quang Hùng CBHD2:

Email1: nqhung@hcmut.edu.vn Email2:

Đào tạo: ☒Chính quy ☐Chất lượng cao

Số lượng sinh viên thực hiện: 5

Mô tả đề tài:

- Giáo dục 4.0 và chuyển đổi số giáo dục yêu cầu kết hợp từ LMS như Moodle, các phần mềm hỗ trợ học trực tuyến trên các thiết bị thông minh (smart devices) như smartphones, tablets, Internet Tivi,…. Hệ thống E-Learning như ongvanghoctap.edu.vn có nguồn học liệu mở, bài tập tương tác, Quiz, Interactive videos, Video meeting, Virtual Lab, v.v..

Đề tài phát triển kiến trúc phần lõi thu gom dữ liệu của người học, các hoạt động trên hệ thống Learning Mangament System (LMS) như https://ongvanghoctap.edu.vn/. Xây dựng các đánh giá hiệu quả của người học trên các khóa học và dự đoán kết quả của người học (dựa trên lịch sử), Cung cấp cho nhà quản lý nhà trường & khoa góc nhìn về toàn cục. Phần này độc lập với các hệ thống LMS.

- Có sử dụng kiến thức về dữ liệu lớn, trí tuệ nhân tạo, Data Lake.

Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu về LMS như Moodle.

- Tìm hiểu về Kafka, Spark

- Hệ thống thu thập dữ liệu từ các LMS khác dựa trên mô hình Data Lake.

Giúp giáo viên so sánh sự tham gia và kết quả với các khóa học khác.

Cung cấp thông tin để giáo viên có thể thực hiện như liên hệ trực tiếp với học sinh, chỉ định nội dung đặc biệt, khuyến khích tham gia hoặc cung cấp dịch vụ dạy kèm đặc biệt.

Giúp HS/SV có thể tìm hiểu về hiệu suất của mình, cá nhân và so sánh với nhóm.

Kết quả cần đạt được:

Đề cương đề xuất mô hình thu gom dữ liệu, mô hình đánh giá người học, phát hiện học sinh trung bình/yếu, dự đoán kết quả học tập của người học.

Tài liệu tham khảo:

Liên hệ GVHD: nqhung@hcmut.edu.vn

Tên đề tài (MR206): Mô phỏng giải thuật cấp phát tài nguyên cho các ứng dụng tính toán tối ưu năng lượng và chất lượng dịch vụ (QoS) dùng SimGrid

Số lượng sinh viên thực hiện: 5

SimGrid - Cài đặt và sử dụng SimGrid mô phỏng giải thuật cấp phát tài nguyên trên Cloud Computing. Tối ưu năng lượng tiêu thụ cho các ứng dụng tính toán. Link: https://simgrid.org/usages.html