**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

---

BÁO CÁO THỰC NGHIỆM

HỌC PHẦN: TÍNH TOÁN HIỆU NĂNG CAO

# ĐỀ TÀI: SỬ DỤNG GOOGLE COLAB XÂY DỰNG, HUẤN LUYỆN MÔ HÌNH NHẬN DẠNG/ PHÂN LOẠI MẶT HÀNG THỰC PHẨM VÀ ĐỒ UỐNG SỬ DỤNG MẠNG NƠ-RON HỌC SÂU CNN

**NHÓM: 9**


**CBHD:** ***TS. Hà Mạnh Đào***

**Sinh viên:**

1. **Trần Quốc Toàn** **2022601265** **(Nhóm trưởng)**
2. **Hoàng Anh Tuấn** **2022601957**
3. **Trịnh Bảo Long** **2022601773**


Hà Nội – Năm 2024

# LỜI MỞ ĐẦU

Trong thời đại công nghệ 4.0, trí tuệ nhân tạo (AI) đã trở thành một công cụ quan trọng, góp phần thay đổi cách con người làm việc và tương tác với thế giới. Nhận diện hình ảnh, một nhánh quan trọng của AI, đang được ứng dụng rộng rãi trong nhiều lĩnh vực như thương mại điện tử, y tế, quản lý kho, và logistics. Với sự hỗ trợ từ các nền tảng mạnh mẽ như Google Colab, việc xây dựng và huấn luyện các mô hình AI không còn là thách thức lớn về tài nguyên và chi phí như trước đây.

Trong cuộc sống hàng ngày, nhu cầu tự động hóa việc phân loại và nhận diện các loại thực phẩm và đồ uống như bánh mì, bia, nước ngọt, hoa quả, rau ngày càng tăng cao. Điều này không chỉ giúp tiết kiệm thời gian mà còn nâng cao độ chính xác trong quản lý hàng hóa, đặc biệt trong các lĩnh vực thương mại bán lẻ và chuỗi cung ứng.

Đề tài "Sử dụng Google Colab xây dựng, huấn luyện mô hình nhận dạng/ phân loại mặt hàng thực phẩm và đồ uống sử dụng mạng nơ-ron học sâu CNN" được thực hiện với mục tiêu nghiên cứu và phát triển một giải pháp AI có thể ứng dụng thực tiễn. Đề tài không chỉ tận dụng sức mạnh của các mô hình học sâu (deep learning) mà còn khai thác hiệu quả nền tảng Google Colab - một công cụ tính toán mạnh mẽ, dễ tiếp cận và miễn phí.

Với việc tập trung vào quy trình từ thu thập dữ liệu, xử lý, xây dựng và đánh giá mô hình, nghiên cứu này không chỉ mang ý nghĩa khoa học mà còn có tính thực tiễn cao. Kết quả của đề tài sẽ đóng góp tích cực vào việc ứng dụng AI trong đời sống và mở ra cơ hội mới cho các nghiên cứu hoặc sản phẩm công nghệ trong tương lai.

# MỤC LỤC

# PHẦN I: PHẦN MỞ ĐẦU

## 1.1. Lý do chọn đề tài

Trong bối cảnh công nghệ 4.0, trí tuệ nhân tạo (AI) đang được ứng dụng rộng rãi trong các lĩnh vực khác nhau, bao gồm nhận diện hình ảnh. Google Colab là một nền tảng mạnh mẽ và miễn phí, hỗ trợ phát triển các mô hình AI với tài nguyên tính toán GPU. Chúng em lựa chọn đề tài "Sử dụng Google Colab xây dựng, huấn luyện mô hình nhận dạng/ phân loại mặt hàng thực phẩm và đồ uống sử dụng mạng nơ-ron học sâu CNN":

- Đáp ứng nhu cầu thực tế trong việc tự động hóa nhận diện và phân loại hình ảnh các mặt hàng phổ biến như bánh mì, bia, nước ngọt, hoa quả, rau.

- Góp phần xây dựng ứng dụng AI hỗ trợ các ngành thương mại, bán lẻ, quản lý kho, và kiểm kê hàng hóa.

- Giúp tận dụng hiệu quả nền tảng Google Colab, giảm thiểu chi phí và thời gian cho việc xây dựng và huấn luyện mô hình AI.

## 1.2. Mục đích nghiên cứu

Xây dựng và huấn luyện một mô hình học sâu (deep learning) có khả năng nhận diện chính xác các loại thực phẩm và đồ uống như bánh mì, bia nước ngọt, hoa quả, rau.

Tích hợp quy trình từ thu thập dữ liệu, tiền xử lý, huấn luyện mô hình đến đánh giá kết quả trong môi trường Google Colab.

Tạo tiền đề cho các nghiên cứu hoặc ứng dụng tiếp theo trong lĩnh vực nhận diện hình ảnh và quản lý hàng hóa.

## 1.3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

- Các kỹ thuật học sâu liên quan đến nhận diện hình ảnh, đặc biệt là sử dụng mạng neural tích chập (Convolutional Neural Networks - CNN).

- Nền tảng Google Colab và các thư viện phổ biến như TensorFlow, PyTorch, hoặc Keras.

Phạm vi nghiên cứu:

- Tập trung vào nhận diện các danh mục cụ thể: bánh mì, bia nước ngọt, hoa quả, rau.

- Phạm vi dữ liệu: sử dụng tập dữ liệu hình ảnh có sẵn hoặc tự xây dựng với quy mô nhỏ để huấn luyện và kiểm tra mô hình.

## 1.4. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học:

- Cung cấp kiến thức và thực nghiệm về ứng dụng AI trong nhận diện hình ảnh.

- Là tài liệu tham khảo cho các nghiên cứu khác trong lĩnh vực học sâu và trí tuệ nhân tạo.

Ý nghĩa thực tiễn:

- Hỗ trợ các doanh nghiệp thương mại, bán lẻ và logistics trong việc tự động hóa quản lý hàng hóa.

- Mở rộng khả năng ứng dụng của AI vào đời sống, đặc biệt trong việc phân loại thực phẩm và đồ uống, giúp tối ưu hóa quy trình kiểm kê và giảm thiểu sai sót thủ công.

## 1.5. Các kỹ năng cần thiết để thực hiện đề tài nghiên cứu

- Kỹ năng tổng hợp kiến thức: tổng hợp các kiến thức đã học được từ trên lớp

- cũng như trên mạng để hoàn thiện bài tập lớn.

- Khả năng đọc hiểu các tài liệu bằng tiếng Anh

- Kỹ năng hoạt động nhóm: phân chia công việc phù hợp, thảo luận nhóm để hoàn thành công việc

- Kỹ năng viết báo cáo

# PHẦN II: KIẾN THỨC CƠ SỞ

## 2.1. Tính toán hiệu năng cao

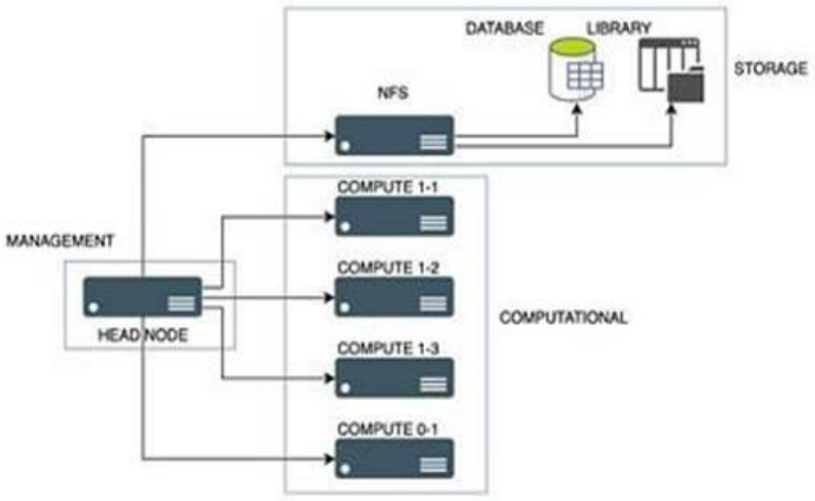
Tính toán hiệu năng cao (HPC) là một lĩnh vực công nghệ thông tin tập trung vào việc sử dụng các hệ thống máy tính mạnh mẽ để thực hiện các phép toán phức tạp và xử lý dữ liệu với tốc độ cao. HPC cho phép các tổ chức giải quyết những vấn đề lớn mà máy tính thông thường không thể xử lý hiệu quả, nhờ vào khả năng tính toán song song và kết nối mạng mạnh mẽ.

2.1.1. Khái niệm:

- Là việc sử dụng tổng hợp sức mạnh của máy tính để mang lại kết quả cao hơn so với máy tính truyền thống hoặc các máy chủ trong việc xử lý các bài toán khó và phức tạp.

- Mô hình hệ thống tính toán hiệu năng cao là một hệ thống các máy tính có kết nối với nhau qua mạng internet hoặc là 1 siêu máy tính được thiết kế để giải quyết các bài toán lớn với tốc độ cao.

- Tính toán hiệu năng cao thường được sử dụng trong một số vấn đề lớn của thế giới về khoa học, kĩ thuật, tài chính, môi trường …




2.1.2. Cấu trúc thành phần

HPC bao gồm ba thành phần chính:

- **Tính toán (Compute)**: Các máy chủ, hay còn gọi là nút (node), làm việc cùng nhau để xử lý các tác vụ. Một cụm HPC có thể bao gồm hàng trăm hoặc hàng ngàn nút, mỗi nút có thể là một máy tính cá nhân hoặc siêu máy tính.

- **Mạng (Network)**: Kết nối giữa các nút trong cụm, cho phép truyền tải dữ liệu nhanh chóng và hiệu quả.

- **Lưu trữ (Storage)**: Hệ thống lưu trữ dữ liệu cần thiết cho các tác vụ tính toán, đảm bảo rằng dữ liệu có thể được truy cập nhanh chóng khi cần thiết

Để xây dựng kiến trúc điện toán hiệu năng cao, các máy chủ tính toán được nối với nhau thành cụm. Chương trình và thuật toán phần mềm được chạy đồng thời trên các máy chủ trong cụm. Các cụm được nối mạng để lưu trữ dữ liệu và nắm bắt đầu ra. Các thành phần này hoạt động liền mạch để hoàn thành một loạt các nhiệm vụ.

Để đạt được hiệu suất tối đa, mỗi thành phần phải theo kịp các thành phần khác. Một ví dụ dễ hiểu, thành phần lưu trữ phải có khả năng cung cấp và nhập dữ liệu đến và từ các máy chủ tính toán ngay khi được xử lý. Cũng như vậy, các thành phần mạng phải có khả năng hỗ trợ vận chuyển dữ liệu tốc độ cao giữa các máy chủ tính toán và lưu trữ dữ liệu. Nếu một thành phần không thể theo kịp phần còn lại, hiệu suất của toàn bộ cơ sở hạ tầng HPC sẽ bị ảnh hưởng.

### 2.1.3. Cách thức hoạt động

HPC hoạt động bằng cách chia nhỏ các tác vụ lớn thành nhiều phần nhỏ hơn, sau đó phân phối chúng cho các nút khác nhau trong cụm để xử lý đồng thời. Điều này giúp tăng tốc độ xử lý tổng thể và giảm thời gian hoàn thành công việc. Để đạt được hiệu suất tối đa, tất cả các thành phần của hệ thống phải hoạt động đồng bộ với nhau.

### 2.1.4. Ứng dụng của HPC

Được triển khai tại các cơ sở hạ tầng, điện toán edge hoặc trên đám mây, các giải pháp HPC được sử dụng cho nhiều mục đích khác nhau:

- Phòng thí nghiệm nghiên cứu: HPC giúp các nhà khoa học tìm ra các nguồn năng lượng tái tạo, hiểu được sự phát triển của vũ trụ, dự đoán, theo dõi các cơn bão và tạo ra các vật liệu mới.

- Truyền thông, giải trí: HPC được sử dụng để chỉnh sửa phim, tạo hiệu ứng đặc biệt, phát trực tiếp các sự kiện trên khắp thế giới.

- Dầu khí: HPC xác định chính xác hơn nơi khoan cho các giếng mới và để giúp thúc đẩy sản xuất từ các giếng hiện có.

- Trí tuệ nhân tạo và máy học: HPC được sử dụng để phát hiện gian lận thẻ tín dụng, cung cấp hỗ trợ kỹ thuật tự hướng dẫn, dạy phương tiện tự lái.

- Các dịch vụ tài chính: HPC được sử dụng để theo dõi xu hướng chứng khoán và tự động hóa giao dịch.

- Sản xuất: HPC được sử dụng để thiết kế các sản phẩm mới, mô phỏng các kịch bản thử nghiệm và đảm bảo rằng các bộ phận được giữ trong kho để dây chuyền sản xuất được ổn định.

- Y tế: HPC được sử dụng để giúp phát triển các phương pháp chữa trị các bệnh như tiểu đường và ung thư và cho phép chẩn đoán bệnh nhân nhanh hơn, chính xác hơn

2.1.5. Lợi ích và vai trò của HPC

HPC mang lại nhiều lợi ích đáng kể, bao gồm:

- **Tăng tốc độ xử lý**: Giúp thực hiện các phép toán lớn trong vài giây thay vì hàng tuần hoặc hàng tháng.

- **Giảm chi phí**: Tăng cường hiệu quả làm việc dẫn đến tiết kiệm thời gian và chi phí cho các dự án nghiên cứu.

- **Giảm nhu cầu kiểm tra vật lý**: Thay vì thử nghiệm thực tế, HPC có thể mô phỏng để kiểm tra các thiết kế

- HPC không chỉ là một công nghệ tiên tiến mà còn là một yếu tố quan trọng trong việc thúc đẩy sự đổi mới và phát triển trong nhiều lĩnh vực khác nhau.

Vai trò của HPC: HPC đã đóng một vai trò quan trọng trong thế giới nghiên cứu với những bước đột phá đáng kể.

- Tốc độ cao: HPC được thiết kế để xử lý tốc độ nhanh có nghĩa là các hệ thống HPC thực hiện các phép tính lớn trong vài giây mà có thể mất hàng tuần hoặc hàng tháng nếu dùng bộ xử lý thông thường.

- Giảm chi phí: Vì tốc độ xử lý cao, các ứng dụng chạy nhanh hơn và do đó các câu trả lời cho các phép tính phức tạp được đưa ra nhanh chóng dẫn đến việc tốn ít tiền và thời gian hơn cho các công việc.

- Giảm kiểm tra vật lý: các ứng dụng hiện đại thường yêu cầu kiểm tra vật lý trước khi sử dụng trên thị trường. Hệ thống HPC có thể tạo mô phỏng nên loại bỏ nhu cầu kiểm tra vật lý, vốn có thể tốn kém và dễ xảy ra lỗi.

Máy tính hiệu suất cao đánh dấu kỷ nguyên đổi mới, nơi máy tính kỹ thuật số sẵn sàng giải quyết các vấn đề lớn. Chỉ là vấn đề thời gian trước khi HPC trở thành xu hướng chủ đạo để mở khóa sức mạnh.

## 2.2. Tính toán song song là gì?

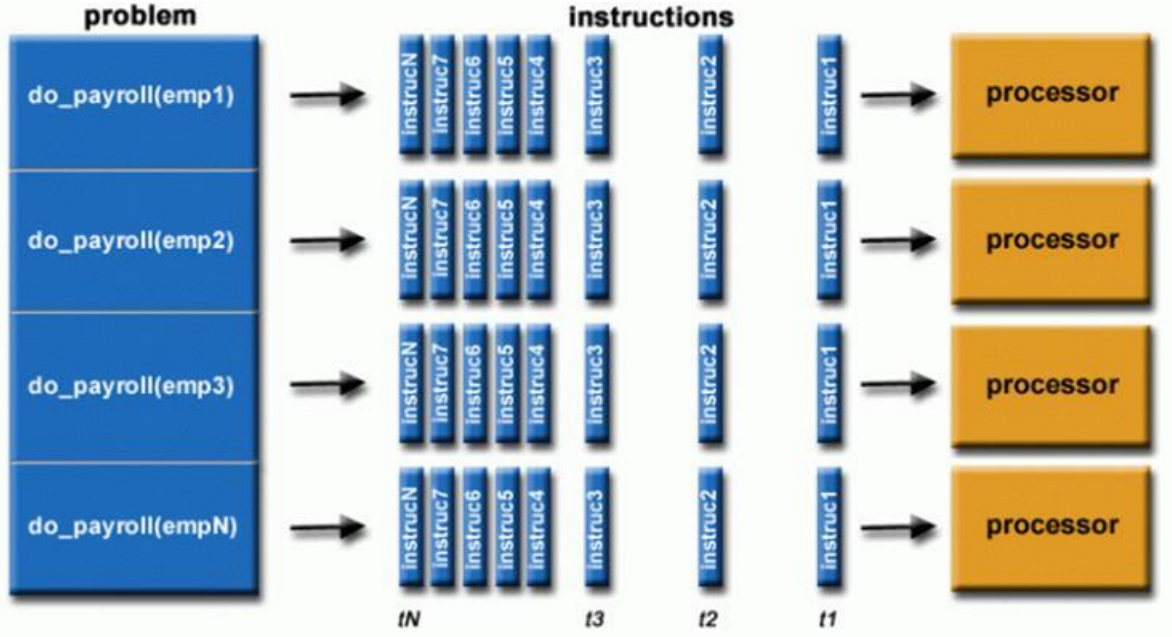
Theo nghĩa đơn giản nhất, tính toán song song là việc sử dụng đồng thời nhiều tài nguyên tính toán để giải quyết một vấn đề tính toán:

- Một vấn đề được chia thành các phần riêng biệt có thể được giải quyết đồng thời.

- Mỗi phần được chia nhỏ thành một loạt các hướng dẫn.

- Các hướng dẫn từ mỗi phần thực thi đồng thời trên các bộ xử lý khác nhau.

- Một cơ chế kiểm soát/phối hợp tổng thể được sử dụng.




Vấn đề tính toán sẽ có thể:

- Chia nhỏ các công việc rời rạc để có thể giải quyết đồng thời;
- Thực hiện nhiều hướng dẫn chương trình bất cứ lúc nào;
- Được giải quyết trong thời gian ngắn hơn với nhiều tài nguyên điện toán so với một tài nguyên điện toán duy nhất.

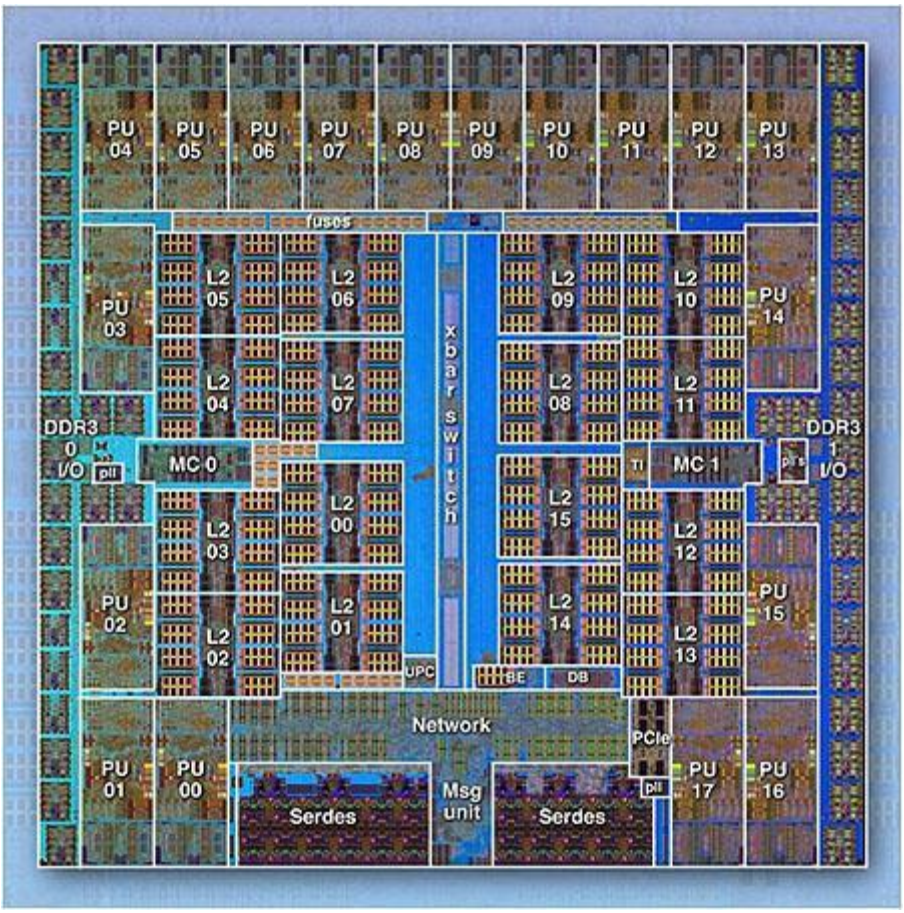
Các tài nguyên tính toán thường là:

- Một máy tính có nhiều bộ xử lý/lõi
- Một số lượng tùy ý các máy tính như vậy được kết nối bởi một mạng

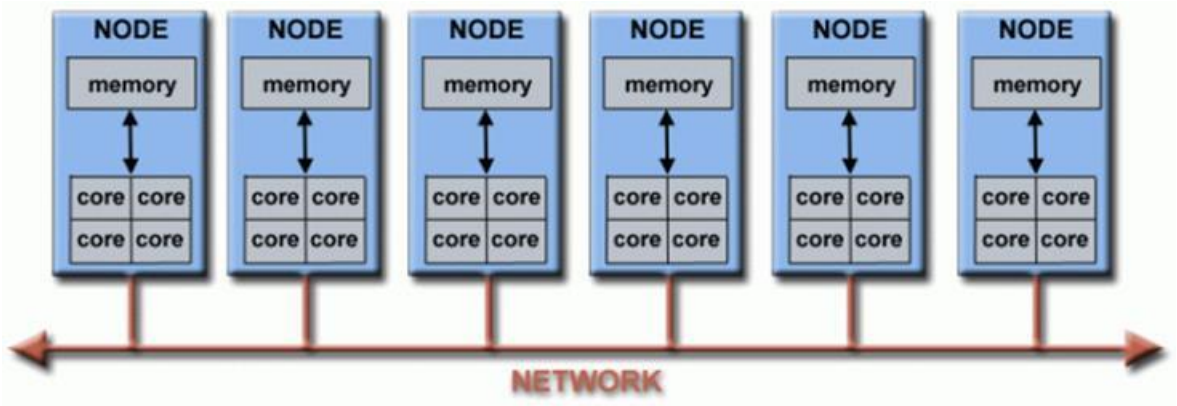
**Máy tính song song:**

Hầu như tất cả các máy tính độc lập ngày nay đều song song từ góc độ phần cứng:

- Nhiều đơn vị chức năng (bộ đệm L1, bộ đệm L2, nhánh, tìm nạp trước, giải mã, dấu phẩy động, xử lý đồ họa (GPU), số nguyên, v.v.)
- Nhiều đơn vị thực thi/lõi
- Nhiều chủ đề phần cứng



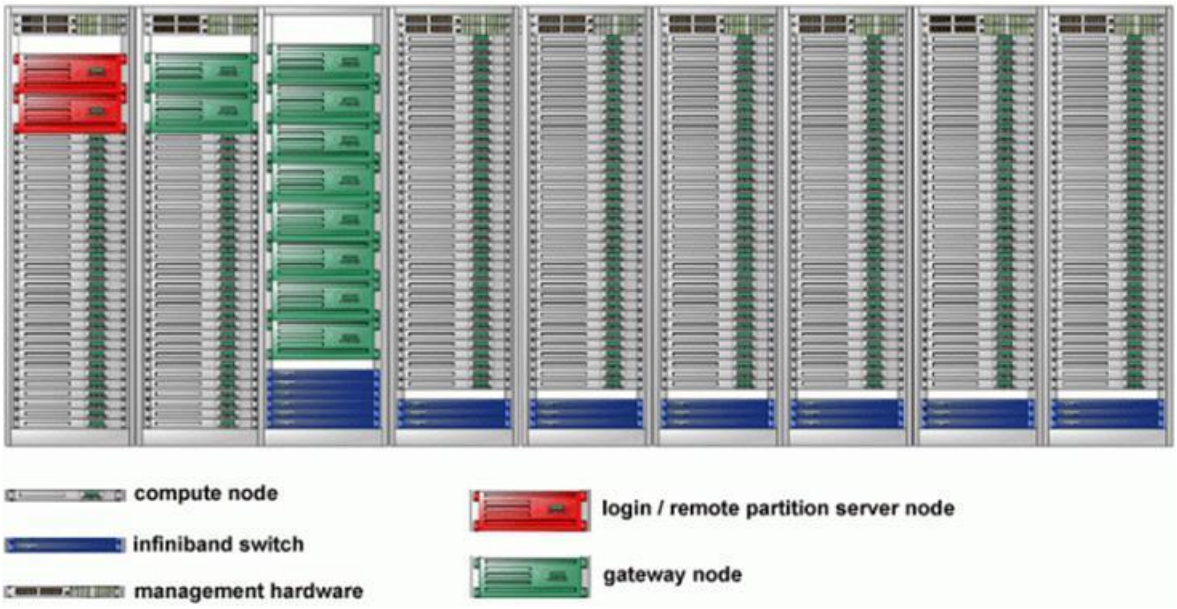

- Các mạng kết nối nhiều máy tính độc lập (các nút) để tạo thành các cụm máy tính song song lớn hơn.




Ví dụ, sơ đồ bên dưới hiển thị một cụm máy tính song song LLNL điển hình:

- Bản thân mỗi nút tính toán là một máy tính song song đa bộ xử lý
- Nhiều nút điện toán được nối mạng cùng với mạng Infiniband

- Các nút mục đích đặc biệt, cũng là bộ xử lý đa năng, được sử dụng cho các mục đích khác




- Phần lớn các máy tính song song lớn (siêu máy tính) trên thế giới là các cụm phần cứng được sản xuất bởi một số ít (hầu hết) các nhà cung cấp nổi tiếng.

## 2.3. Tính toán hiệu năng cao trên nền tảng đám mây

### 2.3.1. Định nghĩa HPC trên Đám Mây

Tính toán hiệu năng cao (High-Performance Computing - HPC) trên nền tảng đám mây (Cloud Computing) là một giải pháp cho phép người dùng sử dụng các tài nguyên tính toán mạnh mẽ thông qua các dịch vụ đám mây thay vì phải đầu tư vào phần cứng HPC truyền thống. Điều này kết hợp giữa sức mạnh tính toán của HPC với sự linh hoạt, tiết kiệm chi phí và khả năng mở rộng của đám mây. HPC trên đám mây kết hợp sức mạnh tính toán cao với cơ sở hạ tầng đám mây, cho phép người dùng truy cập vào các tài nguyên tính toán lớn mà không cần đầu tư vào phần cứng đắt tiền. Các nhà cung cấp dịch vụ đám mây như AWS, Google Cloud và Microsoft Azure cung cấp các giải pháp HPC với khả năng mở rộng linh hoạt và tính toán theo yêu cầu.

Trước đây, các hệ thống HPC bị giới hạn ở dung lượng và thiết kế mà cơ sở hạ tầng tại chỗ có thể cung cấp. Ngày nay, đám mây mở rộng dung lượng cục bộ với các tài nguyên khác.

Nền tảng quản lý đám mây mới nhất cho phép áp dụng phương pháp tiếp cận đám mây kết hợp, kết hợp cơ sở hạ tầng tại chỗ với các dịch vụ đám mây công cộng để khối lượng công việc có thể luân chuyển liền mạch trên tất cả các tài nguyên sẵn có. Điều này cho phép linh hoạt hơn trong việc triển khai các hệ thống HPC và chúng có thể nâng cấp nhanh như thế nào, cùng với cơ hội tối ưu hóa tổng chi phí sở hữu (TCO).

Thông thường, hệ thống HPC tại chỗ cung cấp TCO thấp hơn hệ thống HPC tương đương được dành riêng 24/7 trên đám mây. Tuy nhiên, giải pháp tại chỗ có quy mô phù hợp với công suất tối đa sẽ chỉ được sử dụng hết khi đạt đến công suất tối đa đó. Trong phần lớn thời gian, giải pháp tại chỗ có thể không được sử dụng đúng mức, dẫn đến tài nguyên nhàn rỗi. Tuy nhiên, khối lượng công việc không thể tính toán được do thiếu năng lực sẵn có có thể dẫn đến mất cơ hội.

Nói tóm lại, việc sử dụng đám mây để tăng cường cơ sở hạ tầng HPC tại chỗ của bạn cho các công việc nhạy cảm về thời gian có thể giảm thiểu nguy cơ bỏ lỡ các cơ hội lớn.

2.3.2. Cơ sở hạ tầng

HPC truyền thống: Thường bao gồm các hệ thống tính toán mạnh mẽ với phần cứng được tối ưu hóa cho các tác vụ đòi hỏi tính toán cao, như siêu máy tính, các cluster chuyên dụng.

HPC trên đám mây: Các nhà cung cấp dịch vụ đám mây như AWS, Microsoft Azure, Google Cloud, và IBM Cloud cung cấp các tài nguyên đám mây (máy chủ, lưu trữ, mạng) có thể tùy chỉnh theo nhu cầu sử dụng HPC. Người dùng có thể chọn cấu hình phần cứng ảo (vCPU, GPU, FPGA) theo yêu cầu của tác vụ.

### 2.3.3. Ưu điểm và lợi ích của HPC trên đám mây

#### 2.3.3.1. Ưu điểm

Khả năng mở rộng: Đám mây cho phép mở rộng tài nguyên nhanh chóng, linh hoạt đáp ứng nhu cầu tính toán mà không cần mua thêm phần cứng vật lý.
Tiết kiệm chi phí: Không cần đầu tư lớn vào cơ sở hạ tầng phần cứng, người dùng có thể thuê tài nguyên theo nhu cầu và chỉ trả phí khi sử dụng.

Dễ dàng triển khai: Các nhà cung cấp đám mây cung cấp các dịch vụ quản lý sẵn có, giúp người dùng dễ dàng triển khai các hệ thống HPC mà không cần nhiều kiến thức quản lý hạ tầng.

Truy cập từ xa: Người dùng có thể truy cập tài nguyên HPC từ bất cứ đâu thông qua internet, không bị giới hạn bởi vị trí địa lý của phần cứng.

#### 2.3.3.2. Lợi ích

Lợi ích của HPC trên Đám Mây:

- **Tính linh hoạt và khả năng mở rộng**: Người dùng có thể dễ dàng tăng hoặc giảm quy mô tài nguyên theo nhu cầu, cho phép thực hiện các tác vụ tính toán lớn mà không cần phải duy trì cơ sở hạ tầng vật lý cố định

- **Tiết kiệm chi phí**: Mô hình thanh toán theo mức sử dụng (pay-as-you-go) giúp giảm thiểu chi phí đầu tư ban đầu cho phần cứng. Người dùng chỉ phải trả cho tài nguyên mà họ thực sự sử dụng, giúp tối ưu hóa ngân sách

- **Quản lý tài nguyên hiệu quả**: Các nền tảng HPC đám mây thường cung cấp các công cụ quản lý tự động, cho phép phân bổ và thu hồi tài nguyên một cách linh hoạt, tối đa hóa hiệu suất và giảm thiểu thời gian tài nguyên không hoạt động

- **Khả năng tiếp cận công nghệ tiên tiến**: HPC trên đám mây cho phép người dùng truy cập vào các công nghệ mới nhất như GPU, bộ nhớ cao và lưu trữ hiệu suất cao mà không cần phải nâng cấp phần cứng tại chỗ

### 2.3.4. Thách thức của HPC trên đám mây

Trong khi đám mây HPC (đám mây điện toán hiệu suất cao) mang lại nhiều lợi thế, nó cũng có những thách thức cản trở việc triển khai thành công. Một số thách thức bao gồm:

- Hiệu suất thay đổi: Chia sẻ tài nguyên đám mây có thể tác động tiêu cực đến hiệu suất ứng dụng, đặc biệt là đối với khối lượng công việc điện toán hiệu suất cao, làm gián đoạn khả năng dự đoán và tính nhất quán của hiệu suất ứng dụng.

- Độ trễ mạng và khả năng kết nối: Kết nối mạng nhanh và đáng tin cậy là rất quan trọng đối với các ứng dụng điện toán hiệu suất cao vì độ trễ có thể làm gián đoạn tốc độ và khả năng phản hồi của ứng dụng khi chia sẻ tài nguyên đám mây.

- Độ phức tạp khi truyền dữ liệu: Việc truyền khối lượng dữ liệu lớn đến và đi từ đám mây có thể tốn thời gian và chi phí, đặc biệt là đối với các tập dữ liệu lớn. Các nút thắt cổ chai khi truyền dữ liệu có thể cản trở việc sử dụng hiệu quả các tài nguyên đám mây, ảnh hưởng đến hiệu suất chung.

- Bảo mật và Quyền riêng tư dữ liệu: Việc lưu trữ dữ liệu nhạy cảm hoặc độc quyền trong môi trường đám mây dùng chung làm dấy lên mối lo ngại về bảo mật và tuân thủ. Đảm bảo các biện pháp bảo mật và quyền riêng tư dữ liệu mạnh mẽ trở nên cần thiết để bảo vệ thông tin nhạy cảm.

- Thách thức cấp phép phần mềm: Các ứng dụng HPC thường dựa vào phần mềm và giấy phép chuyên dụng. Quản lý giấy phép phần mềm trong bối cảnh đám mây có thể phức tạp và có khả năng dẫn đến chi phí bổ sung hoặc các vấn đề về tuân thủ.

- Quản lý chi phí hiệu quả: Dịch vụ đám mây cung cấp tính linh hoạt, nhưng mô hình trả tiền khi sử dụng có thể phát sinh chi phí không lường trước nếu việc sử dụng tài nguyên không được giám sát và tối ưu hóa. Việc triển khai các chiến

lược quản lý chi phí hiệu quả là rất quan trọng để ngăn ngừa tình trạng vượt ngân sách.

- Tránh tình trạng khóa nhà cung cấp: Việc di chuyển khối lượng công việc HPC sang hệ sinh thái của một nhà cung cấp đám mây cụ thể có thể dẫn đến tình trạng khóa nhà cung cấp. Điều này hạn chế tính linh hoạt và làm phức tạp quá trình chuyển đổi khối lượng công việc giữa các nhà cung cấp hoặc trở lại các giải pháp tại chỗ.

- Tính di động của dữ liệu giữa các đám mây: Trong các tình huống liên quan đến nhiều nhà cung cấp đám mây hoặc thiết lập đám mây lai, việc di chuyển dữ liệu và khối lượng công việc liền mạch giữa các môi trường đám mây khác nhau có thể phức tạp và đòi hỏi các công cụ và phương pháp chuyên biệt.

- Đảm bảo khả năng tương thích của ứng dụng: Một số ứng dụng HPC được thiết kế để hoạt động trên các kiến trúc phần cứng cụ thể. Đảm bảo khả năng tương thích với các loại phiên bản đám mây và công nghệ ảo hóa có sẵn có thể là mối quan tâm đáng kể.

- Quản lý sự phức tạp: Việc sắp xếp và quản lý khối lượng công việc HPC trên đám mây có thể đòi hỏi các kỹ năng và công cụ chuyên biệt. Việc tích hợp các dịch vụ đám mây với cơ sở hạ tầng và quy trình làm việc HPC hiện có làm tăng tính phức tạp cho quy trình quản lý.

- Rào cản tuân thủ quy định: Các ngành công nghiệp khác nhau có thể có các yêu cầu tuân thủ quy định riêng biệt ảnh hưởng đến việc xử lý và lưu trữ khối lượng công việc HPC. Các yêu cầu này tác động đến việc lựa chọn nhà cung cấp đám mây và chiến lược triển khai.

- Mất quyền kiểm soát cơ sở hạ tầng: Việc chuyển đổi khối lượng công việc HPC lên đám mây đòi hỏi phải từ bỏ một số quyền kiểm soát đối với cơ sở hạ tầng cơ bản. Việc từ bỏ quyền kiểm soát này có thể gây ra lo ngại, đặc biệt là đối với các tổ chức có các điều kiện tiên quyết cụ thể về hiệu suất và bảo mật.

Để vượt qua những thách thức này đòi hỏi phải có kế hoạch tỉ mỉ, thiết kế kiến trúc chu đáo và sử dụng thành thạo các công nghệ và chiến lược phù hợp. Cách tiếp cận này đảm bảo rằng các lợi ích của đám mây HPC có thể được triển khai trong khi giải quyết hiệu quả các nhược điểm tiềm ẩn.

### 2.3.5. Tầm quan trọng của HPC Cloud?

Đám mây HPC (đám mây điện toán hiệu suất cao) quan trọng vì nhiều lý do, vì nó giải quyết được nhiều thách thức và mang lại lợi ích đáng kể cho các tổ chức và nhà nghiên cứu làm việc trên các tác vụ tính toán chuyên sâu. Sau đây là một số lý do tại sao đám mây HPC lại cần thiết:

- Khả năng thích ứng: Cơ sở hạ tầng HPC truyền thống thường có năng lực cố định, hạn chế khả năng thích ứng để xử lý các yêu cầu tính toán đa dạng. Đám mây HPC cho phép mở rộng liền mạch các tài nguyên tính toán để đáp ứng khối lượng công việc đang phát triển, tạo điều kiện quản lý các mô phỏng và phân tích phức tạp và rộng hơn.

- Hiệu quả tài chính: Việc thiết lập và duy trì các cụm HPC chuyên dụng đòi hỏi chi phí ban đầu đáng kể cho phần cứng, phần mềm và cơ sở hạ tầng. Các dịch vụ đám mây HPC tuân thủ mô hình dựa trên mức tiêu thụ, trong đó người dùng chỉ phải chịu chi phí tương xứng với mức sử dụng tài nguyên của họ. Điều này loại bỏ nhu cầu đầu tư ban đầu đáng kể và thúc đẩy tiết kiệm tài chính, đặc biệt là đối với các doanh nghiệp có nhu cầu tính toán thay đổi.

- Tiếp cận toàn cầu: Tài nguyên đám mây HPC có thể truy cập được trên toàn thế giới qua internet. Khả năng truy cập này thúc đẩy sự hợp tác giữa các nhà nghiên cứu và nhóm phân tán về mặt địa lý, hợp lý hóa việc trao đổi dữ liệu, quy trình làm việc và phát hiện.

- Khả năng tùy chỉnh: Nền tảng đám mây HPC cung cấp một loạt các cấu hình phần cứng và môi trường phần mềm, cho phép người dùng tự chủ lựa chọn thiết lập tối ưu phù hợp với nhiệm vụ cụ thể của họ. Khả năng thích ứng này đảm

bảo rằng người dùng có thể chọn tài nguyên phù hợp với khối lượng công việc riêng biệt của họ.

- Kết quả được tăng tốc: Khả năng cung cấp tài nguyên nhanh chóng trong đám mây giúp đẩy nhanh việc bắt đầu các thí nghiệm và mô phỏng cho các nhà nghiên cứu. Điều này dẫn đến thời gian quay vòng nhanh hơn để có kết quả, đẩy nhanh tốc độ nghiên cứu và nỗ lực phát triển.

- Phân bổ tài nguyên hiệu quả: Các nền tảng đám mây HPC thường cung cấp chức năng quản lý và điều phối tài nguyên tự động. Do đó, tài nguyên có thể được phân bổ và hủy phân bổ động khi cần, tối đa hóa việc khai thác tài nguyên trong khi giảm thiểu các trường hợp tài nguyên nhàn rỗi.

- Khả năng phục hồi và sao lưu: Các dịch vụ đám mây HPC thường bao gồm các tính năng như dự phòng và sao chép dữ liệu, đảm bảo an toàn cho dữ liệu và mô phỏng vô giá trước các lỗi phần cứng hoặc gián đoạn không lường trước.

- Hỗ trợ cho các giai đoạn cao điểm: Một số thực thể gặp phải các yêu cầu không thường xuyên hoặc không liên tục đối với điện toán hiệu suất cao. Đám mây HPC cho phép họ truy cập tức thời vào các tài nguyên đám mây trong các giai đoạn cao điểm mà không cần cung cấp cơ sở hạ tầng nội bộ.

- Có thể truy cập cho các thực thể nhỏ hơn: Đám mây HPC dân chủ hóa quyền truy cập vào các tài nguyên điện toán hiệu suất cao. Các tổ chức và nhà nghiên cứu nhỏ hơn không có đủ nguồn lực để đầu tư vào phần cứng HPC chuyên dụng có thể khai thác các dịch vụ đám mây để thực hiện các phép tính nâng cao.

- Thúc đẩy sự đổi mới: Đám mây HPC phá bỏ các rào cản cản trở việc thử nghiệm và khám phá các khái niệm mới, trao quyền cho các nhà nghiên cứu đổi mới và khám phá những hiểu biết mới với hiệu quả cao hơn.

Đám mây HPC mang đến một giải pháp linh hoạt, tiết kiệm chi phí và khả thi cho các tổ chức và nhà nghiên cứu để khai thác khả năng điện toán hiệu suất

cao mà không cần đến sự phức tạp và hạn chế của cơ sở hạ tầng HPC tại chỗ thông thường.

## 2.4. Các dịch vụ HPC tiêu biểu trên đám mây

Các Nhà Cung Cấp Dịch Vụ HPC Trên Đám Mây

- **Amazon Web Services (AWS)**: AWS cung cấp một loạt dịch vụ HPC cho phép người dùng chạy các mô phỏng phức tạp và khối lượng công việc học sâu trên đám mây. Nền tảng này hỗ trợ khả năng tính toán gần như không giới hạn và hệ thống lưu trữ hiệu suất cao, giúp tăng tốc độ đổi mới và tối đa hóa hiệu quả hoạt động.

- **Microsoft Azure**: Azure cung cấp các giải pháp HPC thông qua Azure CycleCloud và Azure Batch, cho phép quản lý khối lượng công việc HPC và tự động tính toán tài nguyên cần thiết. Azure cũng hợp tác với nhiều nhà cung cấp phần cứng để tối ưu hóa hiệu suất cho các ứng dụng HPC.

- **Google Cloud Platform (GCP)**: Google Cloud cung cấp các giải pháp HPC dễ sử dụng, được tối ưu hóa về chi phí, với khả năng triển khai nhanh chóng thông qua các bộ công cụ HPC. GCP cho phép người dùng truy cập vào các tài nguyên tính toán mạnh mẽ và quản lý chi phí hiệu quả khi mở rộng quy mô.

- **Hewlett Packard Enterprise (HPE)**: HPE tích hợp các nguồn lực HPC với hạ tầng đám mây, cung cấp dịch vụ HPC cho phép tổ chức và nhà nghiên cứu tận dụng sức mạnh tính toán mà không cần quản lý các cụm HPC riêng biệt.

- **Oracle**: Oracle cũng cung cấp dịch vụ HPC với hiệu suất cạnh tranh, sử dụng các bộ xử lý Intel® Xeon® Scalable để tối ưu hóa hiệu suất mô phỏng và tính toán trong đám mây.

- **Penguin Computing**: Công ty này cung cấp giải pháp HPC trên đám mây với khả năng mở rộng cao và quản lý đơn giản. Penguin Computing hỗ trợ cả môi trường HPC vật lý và đám mây, giúp khách hàng dễ dàng triển khai các giải pháp HPC.

### 2.4.1. HPC của Google Cloud

Trong bối cảnh HPC đang phát triển nhanh chóng, Google Cloud liên tục thúc đẩy ranh giới của sự đổi mới để cung cấp các giải pháp phù hợp cho khối lượng công việc đòi hỏi khắt khe nhất. Với những bước tiến gần đây trong phát triển sản phẩm HPC, Google Cloud đã giới thiệu một loạt các sản phẩm tùy chỉnh và được xây dựng theo mục đích cụ thể để phục vụ riêng cho khách hàng HPC. Được xây dựng trên cơ sở hạ tầng mới nhất của Google, các giải pháp này dễ sử dụng và được tối ưu hóa về chi phí để cung cấp nền tảng linh hoạt và mạnh mẽ cho điện toán hiệu suất cao.

Google Cloud HPC là nền tảng mạnh mẽ cho các tác vụ đòi hỏi khả năng tính toán cao, bao gồm mô phỏng khoa học, xử lý dữ liệu lớn, AI/ML, và nhiều ứng dụng khác.

Các Giải Pháp HPC Trên Google Cloud

- **H3 Virtual Machines**: GCP mới giới thiệu dòng máy ảo H3, được tối ưu hóa cho các khối lượng công việc HPC. Các máy này sử dụng bộ xử lý Intel Xeon thế hệ thứ 4, cung cấp hiệu suất cao với 88 lõi và 352 GB bộ nhớ, giúp tăng tốc độ xử lý cho nhiều ứng dụng HPC khác nhau

- **Cloud HPC Toolkit**: Đây là một bộ công cụ mã nguồn mở giúp đơn giản hóa quá trình thiết lập môi trường HPC. Người dùng có thể triển khai các cụm siêu máy tính chỉ với ít dòng mã, giúp tiết kiệm thời gian và công sức so với việc xây dựng từ đầu

- **Batch Processing**: GCP cung cấp dịch vụ Batch cho phép người dùng chạy các tác vụ tính toán theo lô trên Google Kubernetes Engine (GKE). Dịch vụ này tự động phân bổ tài nguyên tính toán, giúp tối ưu hóa quy trình làm việc và giảm thời gian chờ đợi

- **Lưu trữ và Quản lý Dữ liệu**: GCP cung cấp nhiều tùy chọn lưu trữ như Google Cloud Storage cho dữ liệu không cấu trúc và Cloud Filestore cho lưu trữ

tệp hiệu suất cao. Điều này rất quan trọng trong môi trường HPC, nơi yêu cầu lưu trữ lớn cho dữ liệu đầu vào và kết quả

2.4.1.1. Cấu trúc thành phần

Xây dựng môi trường điện toán hiệu suất cao (HPC) trên Google Cloud Platform (GCP) bao gồm một số thành phần chính hoạt động cùng nhau để cung cấp cơ sở hạ tầng có khả năng mở rộng, đáng tin cậy và hiệu quả để chạy khối lượng công việc tính toán chuyên sâu. Trong câu trả lời này, chúng ta sẽ khám phá chi tiết các thành phần này, tập trung vào vai trò và tầm quan trọng của chúng trong việc tạo môi trường HPC trên GCP.

1. Máy ảo (VM): VM là khối xây dựng cơ bản của bất kỳ môi trường HPC nào. GCP cung cấp nhiều loại VM, bao gồm các phiên bản có bộ nhớ cao, CPU cao và hỗ trợ GPU, được tối ưu hóa cho các loại khối lượng công việc khác nhau. Các VM này có thể được cung cấp và quản lý bằng dịch vụ Compute Engine của GCP. Khi xây dựng môi trường HPC, điều cần thiết là phải chọn loại VM phù hợp dựa trên các yêu cầu cụ thể của khối lượng công việc.

2. Mạng: Mạng đóng vai trò quan trọng trong môi trường HPC vì nó cho phép giao tiếp giữa các nút tính toán và tài nguyên lưu trữ. GCP cung cấp cơ sở hạ tầng mạng mạnh mẽ bao gồm Virtual Private Cloud (VPC), cho phép bạn tạo môi trường mạng riêng biệt. Ngoài ra, GCP cung cấp các tính năng như cân bằng tải, quy tắc tường lửa và kết nối Virtual Private Network (VPN), rất cần thiết để tạo môi trường HPC an toàn và có thể mở rộng.

3. Lưu trữ: Khối lượng công việc HPC thường yêu cầu dung lượng lưu trữ lớn để lưu trữ dữ liệu đầu vào, kết quả trung gian và dữ liệu đầu ra. GCP cung cấp nhiều tùy chọn lưu trữ khác nhau có thể được tận dụng trong môi trường HPC. Google Cloud Storage cung cấp khả năng lưu trữ đối tượng có thể mở rộng cho dữ liệu phi cấu trúc, trong khi Cloud Filestore cung cấp khả năng lưu trữ tệp hiệu suất cao để chia sẻ quyền truy cập. Đối với khối lượng công việc đòi hỏi khắt khe

hơn, GCP cung cấp các tùy chọn như Cloud Block Storage và Cloud Filestore High Scale, mang lại hiệu suất và thông lượng cao hơn.

4. Quản lý dữ liệu: Quản lý dữ liệu hiệu quả là rất quan trọng trong môi trường HPC. GCP cung cấp một số dịch vụ giúp quản lý dữ liệu hiệu quả. Google Cloud Dataflow cho phép xử lý và chuyển đổi dữ liệu phân tán, trong khi BigQuery cung cấp kho dữ liệu không cần máy chủ được quản lý hoàn toàn để phân tích ad-hoc. Ngoài ra, Dịch vụ truyền dữ liệu của GCP cho phép bạn truyền khối lượng lớn dữ liệu vào và ra khỏi đám mây một cách hiệu quả.

5. Orchestration và Job Scheduling: Để chạy khối lượng công việc HPC phức tạp, cần có một hệ thống orchestration và job schedule. GCP cung cấp một số tùy chọn cho mục đích này. Google Cloud Composer cung cấp dịch vụ orchestration workflow được quản lý hoàn toàn dựa trên Apache Airflow. Ngoài ra, bạn có thể sử dụng các giải pháp như Kubernetes Engine hoặc Cloud Dataflow để lên lịch và thực hiện job.

6. Giám sát và ghi nhật ký: Giám sát và ghi nhật ký rất quan trọng để duy trì hiệu suất và độ tin cậy của môi trường HPC. GCP cung cấp các công cụ như Stackdriver Monitoring và Stackdriver Logging, cho phép bạn giám sát việc sử dụng tài nguyên, theo dõi số liệu hiệu suất và khắc phục sự cố hiệu quả. Các công cụ này có thể được tích hợp với các dịch vụ GCP khác để cung cấp khả năng hiển thị toàn diện vào môi trường HPC.

7. Bảo mật và tuân thủ: Bảo mật là yếu tố quan trọng nhất trong bất kỳ môi trường điện toán nào và HPC cũng không ngoại lệ. GCP cung cấp các tính năng bảo mật mạnh mẽ, bao gồm quản lý danh tính và quyền truy cập (IAM), mã hóa khi lưu trữ và khi truyền tải, và các dịch vụ bảo mật chuyên dụng như Cloud Security Command Center. GCP cũng tuân thủ nhiều tiêu chuẩn và quy định của ngành, khiến nó phù hợp với khối lượng công việc HPC đòi hỏi các yêu cầu bảo mật và tuân thủ nghiêm ngặt.

Xây dựng môi trường HPC trên Google Cloud Platform liên quan đến một số thành phần chính, bao gồm máy ảo, mạng, lưu trữ, quản lý dữ liệu, sắp xếp và lập lịch công việc, giám sát và ghi nhật ký, bảo mật và tuân thủ. Bằng cách tận dụng hiệu quả các thành phần này, các tổ chức có thể tạo ra môi trường HPC có khả năng mở rộng, đáng tin cậy và hiệu quả trên GCP.

2.4.1.2. Tính năng nổi bật:

TPU (Tensor Processing Units): TPU là bộ xử lý do Google phát triển, tối ưu hóa cho việc huấn luyện và triển khai các mô hình học sâu, đặc biệt là với TensorFlow.

GPU (Graphics Processing Units): Google Cloud cung cấp các loại GPU mạnh mẽ như NVIDIA Tesla K80, P100, và V100, phù hợp với các tác vụ tính toán phức tạp và các ứng dụng AI/ML.

Máy ảo HPC chuyên dụng: Google Cloud cung cấp các máy ảo (VM) có cấu hình cao, tối ưu hóa cho các tác vụ HPC với khả năng mở rộng linh hoạt.

Tích hợp Colab: Dành cho các nhà nghiên cứu và sinh viên, Google Colab cung cấp môi trường tính toán miễn phí với GPU/TPU, thuận tiện cho việc huấn luyện các mô hình AI và thử nghiệm.

2.4.2. HPC của Amazon Web Services (AWS)

AWS HPC là nền tảng đám mây tiên tiến giúp các tổ chức thực hiện các tác vụ tính toán khối lượng lớn với cơ sở hạ tầng đám mây mở rộng.

Tính năng nổi bật:

- **Amazon Elastic Compute Cloud (EC2)**: Cung cấp khả năng tính toán linh hoạt với hơn 400 loại phiên bản khác nhau, bao gồm cả CPU và GPU. EC2 cho phép người dùng chọn cấu hình phù hợp với nhu cầu tính toán của họ, từ các ứng dụng HPC quy mô lớn đến các dự án học máy.

- **Elastic Fabric Adapter (EFA)**: Đây là một công nghệ mạng cho phép chạy các ứng dụng HPC quy mô lớn với hiệu suất cao. EFA cung cấp khả năng giao tiếp

giữa các nút với độ trễ thấp, giúp tăng tốc độ xử lý cho các khối lượng công việc phân tán.

- **AWS ParallelCluster**: Một công cụ mã nguồn mở giúp người dùng dễ dàng triển khai và quản lý các cụm HPC trên AWS. Nó tự động hóa quá trình thiết lập môi trường tính toán, tiết kiệm thời gian và công sức cho người dùng.

- **AWS Batch**: Dịch vụ này cho phép người dùng chạy hàng triệu tác vụ tính toán theo lô trên AWS, tự động quản lý và tối ưu hóa tài nguyên cần thiết cho từng tác vụ.

- **Amazon FSx for Lustre**: Cung cấp một hệ thống lưu trữ hiệu suất cao, giúp xử lý nhanh chóng các tập dữ liệu lớn với độ trễ dưới một mili giây, rất cần thiết cho các ứng dụng HPC.

2.4.3. HPC của Microsoft Azure

Microsoft Azure HPC là giải pháp linh hoạt cho các ứng dụng yêu cầu khả năng tính toán khối lượng lớn, với nhiều dịch vụ chuyên biệt dành cho các nhà nghiên cứu và doanh nghiệp.

Tính năng nổi bật:

Azure Virtual Machines (VMs): Cung cấp nhiều loại VM tối ưu hóa cho HPC, bao gồm các loại có GPU mạnh như NVIDIA Tesla V100 hoặc A100.

Azure CycleCloud: Dịch vụ này giúp bạn tự động hóa việc tạo lập, quản lý và tối ưu hóa các cụm HPC trên Azure, hỗ trợ việc triển khai trên quy mô lớn.

InfiniBand: Một mạng tốc độ cao với độ trễ thấp, lý tưởng cho các ứng dụng HPC yêu cầu băng thông lớn, chẳng hạn như các mô phỏng vật lý hoặc phân tích dữ liệu lớn.

AI và Big Data Integration: Azure tích hợp mạnh mẽ các công cụ AI và phân tích dữ liệu lớn, giúp dễ dàng triển khai các ứng dụng HPC kết hợp với học máy và phân tích.

### 2.4.4. Các đám mây Việt Nam cung cấp HPC

Tại Việt Nam, các nhà cung cấp dịch vụ đám mây lớn như Viettel Cloud, FPT Cloud, VNG Cloud cũng bắt đầu cung cấp các giải pháp HPC nhằm đáp ứng nhu cầu trong nước về tính toán hiệu năng cao, đặc biệt là trong các lĩnh vực công nghiệp, tài chính, giáo dục và nghiên cứu.

Các nhà cung cấp nổi bật:

Viettel Cloud: Cung cấp các máy chủ HPC chuyên dụng cho các doanh nghiệp, với các lựa chọn về cấu hình cao, mạng tốc độ lớn, đáp ứng yêu cầu về tính toán và phân tích dữ liệu phức tạp.

FPT Cloud: Có các giải pháp HPC hỗ trợ AI/ML và xử lý dữ liệu lớn, với hạ tầng mạnh mẽ dựa trên điện toán đám mây, giúp các doanh nghiệp trong nước dễ dàng tiếp cận các công nghệ tiên tiến.

VNG Cloud: Hỗ trợ HPC cho các tác vụ như mô phỏng, phân tích dữ liệu, và các ứng dụng công nghiệp yêu cầu năng lực tính toán cao.

### 2.4.5. Các lĩnh vực ứng dụng

HPC trên đám mây thường được sử dụng trong các lĩnh vực đòi hỏi tính toán phức tạp, bao gồm:

- Mô phỏng và dự đoán thời tiết: Xử lý khối lượng dữ liệu lớn và yêu cầu mô hình hóa thời gian thực.

- Hóa học tính toán: Nghiên cứu các phản ứng hóa học và tương tác giữa các phân tử.

- Sinh học tính toán: Phân tích dữ liệu gen, mô phỏng phân tử.

- Tài chính: Các mô hình phân tích rủi ro và dự báo thị trường yêu cầu tính toán nhanh và chính xác.

- Trí tuệ nhân tạo (AI) và học sâu (Deep Learning): Đào tạo các mô hình AI phức tạp đòi hỏi sự tính toán mạnh từ GPU và CPU.

### 2.4.6. Công cụ và phần mềm hỗ trợ

Nhiều công cụ và phần mềm hỗ trợ HPC đã được tối ưu hóa để hoạt động trên đám mây, chẳng hạn như:

- SLURM: Hệ thống quản lý khối lượng công việc HPC.
- MPI (Message Passing Interface): Hỗ trợ tính toán song song giữa các nút trong một hệ thống HPC.
- Docker/Kubernetes: Giúp triển khai và quản lý các ứng dụng HPC trên nền tảng đám mây theo cách dễ dàng và nhất quán.

# PHẦN III: XÂY DỰNG VÀ HUẤN LUYỆN MẠNG NƠ-RON HỌC SÂU

## 3.1. Giới thiệu và cài đặt công cụ Google Colaboratory

### 3.1.1. Giới thiệu về Google Colab

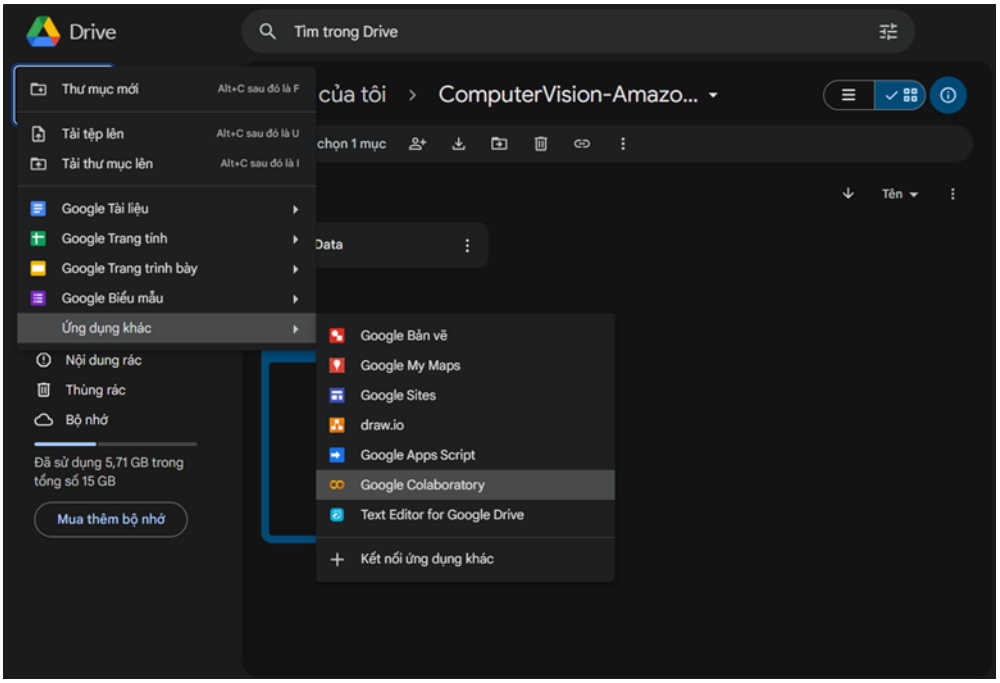
Google Colab (Google Colaboratory) là một dịch vụ đám mây miễn phí của Google nhằm hỗ trợ cộng đồng nghiên cứu AI phát triển các ứng dụng học sâu (deep learning) bằng việc cung cấp GPU và TPU miễn phí.

Google Colab được cài đặt sẵn những thư viện rất phổ biến trong nghiên cứu Deep Learning như PyTorch, TensorFlow, Keras và OpenCV.

### 3.1.2. Cài đặt Google Colab

- Đăng nhập Gmail, truy cập vào Drive

- Kích chọn My Drive/ chọn More/ chọn Connect more apps/ tại ô tìm kiếm gõ Colaboratory/ kích chọn biểu tượng Colaboratory/ chọn Install và cài đặt theo hướng dẫn.




## 3.2. Giới thiệu thư viện Tensorflow

- Tensorflow là thư viện mã nguồn mở hỗ trợ học máy và học sâu nổi tiếng nhất thế giới, được phát triển bởi các nhà nghiên cứu của Google. Việc hỗ trợ

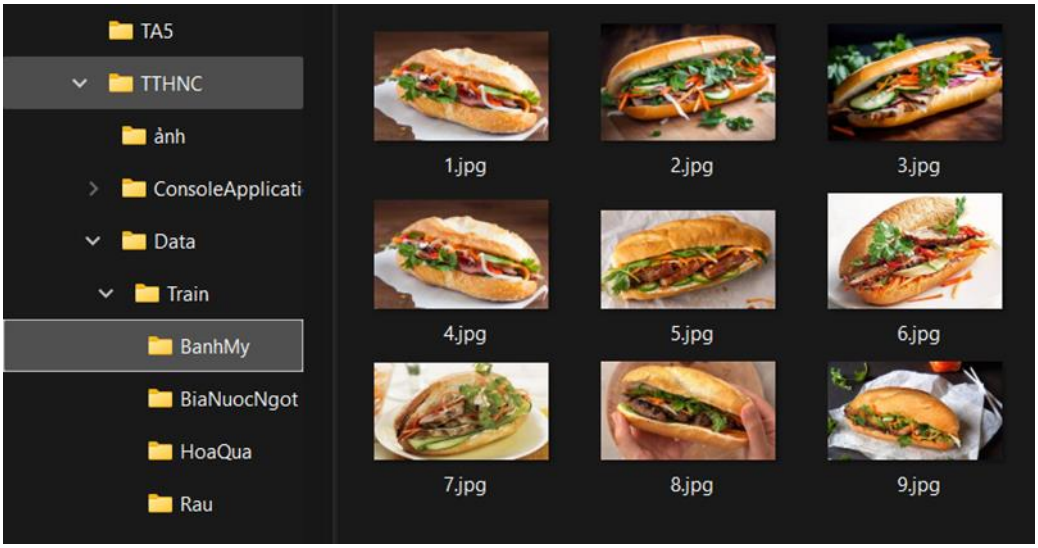
mạnh mẽ các phép toán học để tính toán trong học máy và học sâu đãgiúp việc tiếp cận các bài toán trở nên đơn giản, nhanh chóng và tiện lợi.

- TensorFlow có thể được sử dụng online với Google Colab hoặc cài đặt offline trên máy tính với Anaconda. Nếu máy tính của bạn không có GPU thì nên sửdụng Tensorflow trên Google Colab.

- Kiến trúc TensorFlow hoạt động được chia thành 3 phần: Tiền xử lý dữ liệu, Dựng model, Huấn luyện và sử dụng model

### 3.3. Bài toán ứng dụng

- Cho một tập dữ liệu ảnh về 4 loại sản phẩm Bánh mỳ, Bia nước ngọt, Hoa quả, Rau được bán tại cửa hàng AmazonGo. Xây dựng một mô hình có khả năng nhận dạng được 4 loại sản phẩm này.



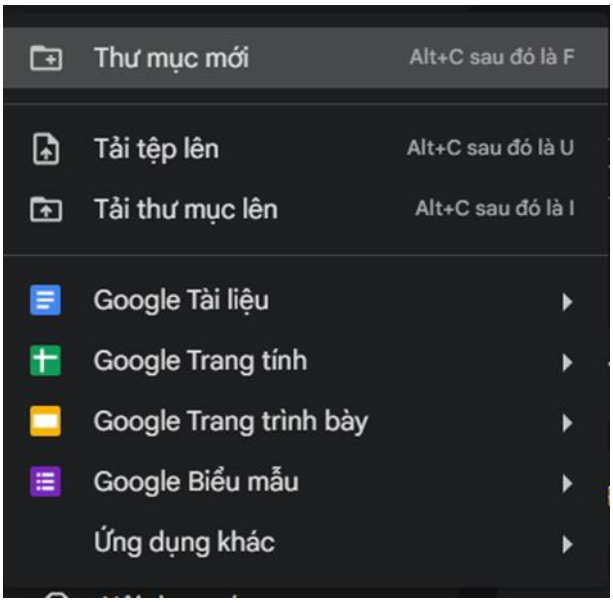

### 3.4. Chuẩn bị dữ liệu

Thu thập dữ liệu ảnh gồm 4 loại sản phẩm Bánh mỳ, Bia nước ngọt, Hoa quả và Rau lưu trong máy tính, giả sử tại thư mục D:\ComputerVision-AmazonGo\Data và hai thư mục con Train và Validation; Mỗi loại sản phẩm tương ứng một thưmục trong đó có 8 ảnh Train, 2 ảnh Validation.

### 3.5. Xây dựng mô hình nhận ảnh sử dụng mạng nơ-ron học sâu CNN trên Google Colab

3.5.1. Tạo thư mục dự án

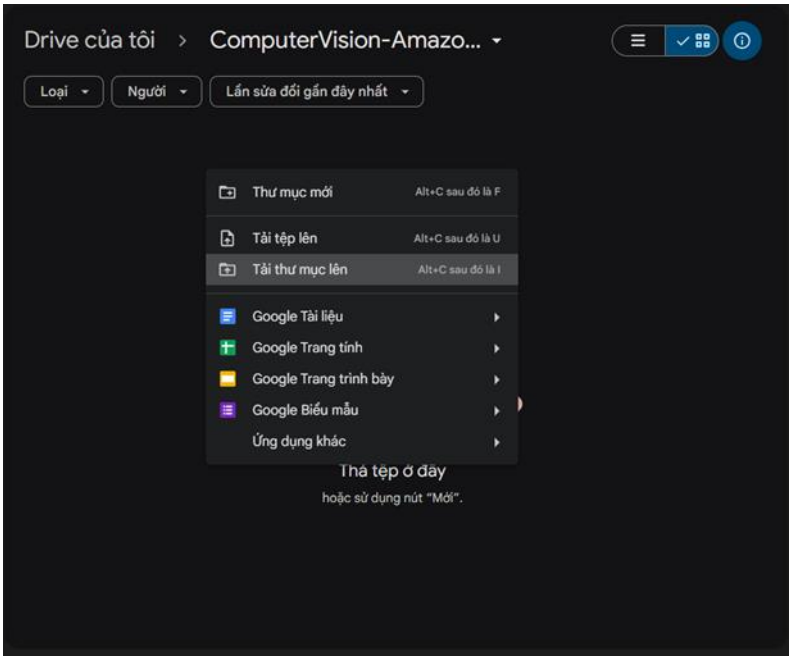
- Đăng nhập Gmail, truy cập vào Drive
- Kích chọn My Drive/ chọn New folder




- Đặt tên thư mục là ComputerVision-AmazonGo/ chọn CREATE.


3.5.2. Upload tập dữ liệu ảnh

- Kích đúp chuột mở thư mục ComputerVision-AmazonGo.
- Kích chuột phải tại vùng trống của thư mục, chọn Upload folder

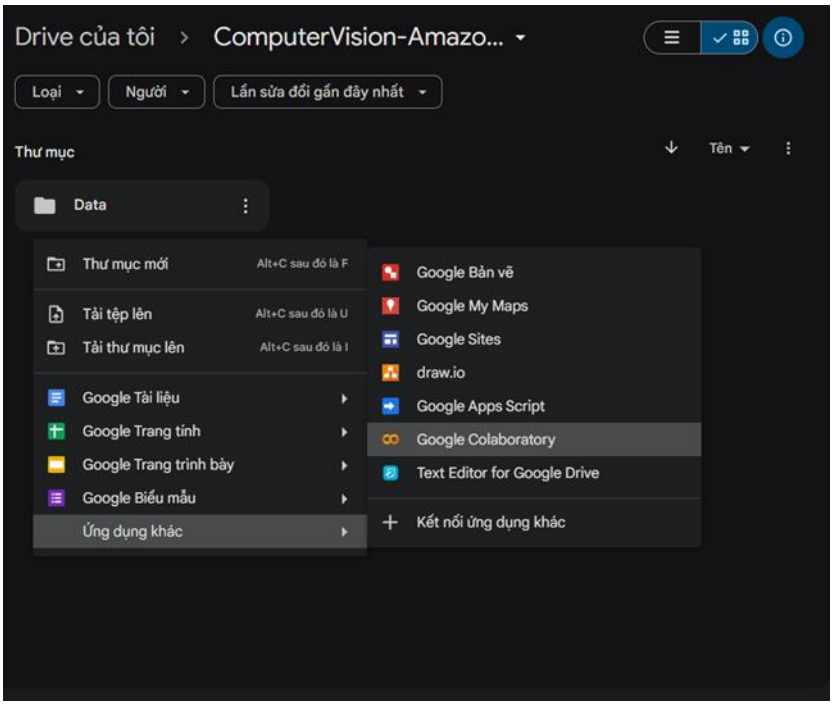



- Chọn đường dẫn tới thư mục Data đã tạo ra ở Bước 2 – Chuẩn bị dữ liệu, chọn thư mục Data, kích chọn Tải lên.

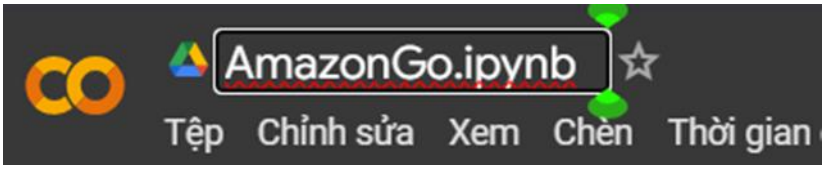
- Kết quả tất cả các thư mục và ảnh chứa trong thư mục Data được tải lên Drive, trong thư mục ComputerVision-AmazonGo

3.5.3. Tạo file Colab notebook trên Google Drive

- Kích chọn My Drive/ chọn More/ chọn Colaboratory.



- Đặt tên cho tệp colab mới là AmazonGo.ipynb




3.5.4. Viết code

3.5.4.1. Khai báo các thư viện sử dụng

```
[1]  import tensorflow as tf
     from tensorflow import keras
     import matplotlib.pyplot as plt
     import numpy as np
```

3.5.4.2. Kết nối với Google Drive để đọc và lưu dữ liệu

```
[2]  from google.colab import drive
     drive.mount('/content/drive')
```

- Chạy đoạn code kết quả hiện ra như sau:

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

3.5.4.3. Khai báo đường dẫn thư mục chứa dữ liệu ảnh

- Khai báo đường dẫn đến thư mục chứa ảnh để Huấn luyện và Kiểm định mô hình

```
[3]  import os
     train_image_files_path = "/content/drive/My Drive/ComputerVision-AmazonGo/Data/Train/"
     valid_image_files_path = "/content/drive/My Drive/ComputerVision-AmazonGo/Data/Validation/"
```

3.5.4.4. Gán nhãn dữ liệu

Phân loại ảnh là bài toán học có giám sát, do đó dữ liệu huấn luyện và kiểm định phải được gán nhãn. Ảnh được gán nhãn dựa theo tên thư mục chứa nó, ta có tên và thứ tự các nhãn tương ứng với tên và thứ tự các thư mục chứa ảnh huấn luyện và kiểm định.

```
[4]  label = ['BanhMy', 'BiaNuocNgot', 'HoaQua', 'Rau']
```

3.5.4.5. Tiền xử lý dữ liệu ảnh với ImageDataGenerator

```
[5]  from tensorflow.keras.preprocessing.image import ImageDataGenerator

     train_data_gen = ImageDataGenerator(rescale=1./255)  # Đọc và chuẩn hóa dữ liệu ảnh về 0-1
     validation_data_gen = ImageDataGenerator(rescale=1./255)
```

- Tham số rescale=1/255 có tác dụng chuẩn hóa dữ liệu ảnh về các giá trị nằm trong khoảng [0, 1].

- Một file ảnh JPEG được lưu trong máy tính dưới dạng một ma trận dữ liệu số có giá trị trong khoảng [0, 255]

Chúng ta nhìn thấy

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 157 | 48 | 240 | 49 | 2 | 78 | 229 | 64 |
| 159 | 54 | 94 | 218 | 126 | 97 | 60 | 163 |
| 128 | 201 | 202 | 100 | 53 | 5 | 4 | 131 |
| 199 | 132 | 202 | 121 | 119 | 238 | 49 | 220 |
| 192 | 63 | 21 | 129 | 16 | 226 | 104 | 32 |
| 124 | 225 | 229 | 180 | 141 | 155 | 153 | 100 |
| 90 | 17 | 238 | 232 | 34 | 209 | 64 | 187 |
| 212 | 244 | 86 | 30 | 192 | 160 | 85 | 195 |
| 226 | 221 | 201 | 223 | 161 | 170 | 114 | 154 |
| 252 | 217 | 1 | 107 | 127 | 126 | 50 | 26 |

Máy tính thấy

- Để biểu diễn một bức ảnh 256x256 pixel trong máy tính thì ta cần ma trận có kích thước 256x256 chiều, mỗi phần tử trong ma trận có giá trị nằm trong khoảng từ 0 đến 255. Tùy thuộc vào bức ảnh là màu hay ảnh xám thì ma trận này sẽ có số kênh tương ứng. Ví dụ với ảnh màu 256x256 RGB, chúng ta sẽ có 3 ma trận 256x256 để biểu diễn ảnh này, với ảnh xám 256x256, chúng ta sẽ có 1 ma trận 256x256 để biểu diễn.

3.5.4.6. Đọc dữ liệu train và validation

```
[6] train_generator = train_data_gen.flow_from_directory(
        train_image_files_path,  # Đường dẫn tới ảnh huấn luyện
        target_size=(50, 50),  # Biến đổi các ảnh huấn luyện về cùng một kích thước [50x50]
        class_mode='categorical'  # Phân loại ảnh đa lớp (4 lớp)
    )
    validation_generator = validation_data_gen.flow_from_directory(
        valid_image_files_path,
        target_size=(50, 50),
        class_mode='categorical'
    )
```

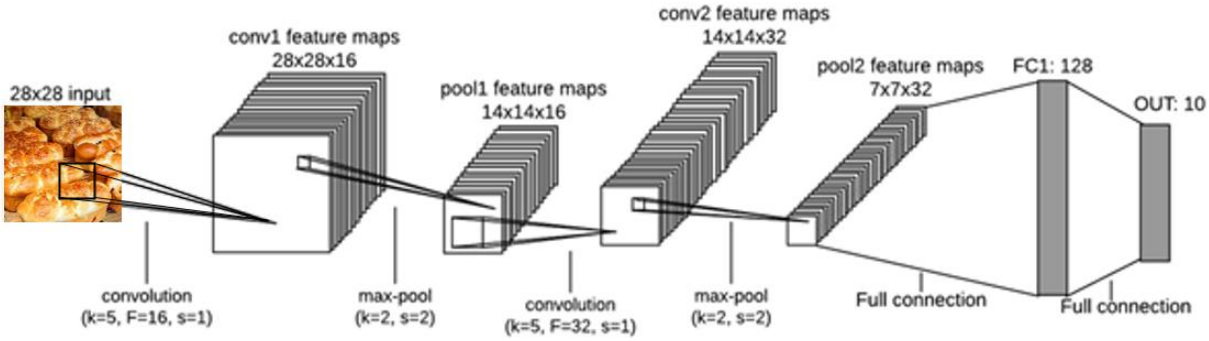
Kết quả chạy ta có:

```
Found 35 images belonging to 4 classes.
Found 4 images belonging to 4 classes.
```

3.5.4.7. Xây dựng mô hình

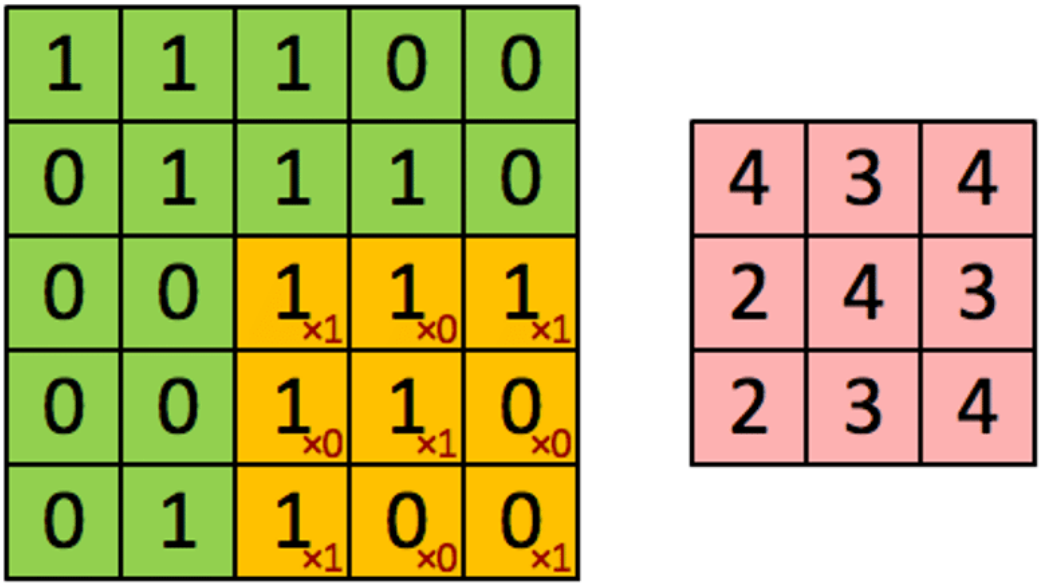
```
[7] from keras.models import Sequential
    from keras.layers import Dense, Dropout, Conv2D, MaxPooling2D, Flatten

    model = tf.keras.models.Sequential()
    #lớp CNN1
    model.add(Conv2D(32, (3,3), activation='relu', input_shape=(50,50,3)))
    model.add(MaxPooling2D(2,2))

    #lớp CNN2
    model.add(Conv2D(64, (3,3), activation='relu'))
    model.add(MaxPooling2D(2,2))

    model.add(Flatten())
    model.add(Dense(512, activation=tf.nn.relu))
    model.add(Dense(4, activation=tf.nn.softmax))
```

- Mô hình gồm 5 tầng: Input image → CNN1 → CNN2 → Fully connected layer → Output.




- Tầng CNN1 gồm 32 bộ lọc kích thước 3x3 (kích thước thường dùng số lẻ 3,5,7). Tầng CNN 1 kết nối với đầu vào nên phải mô tả rõ thông tin của đầu vào (input_shape).

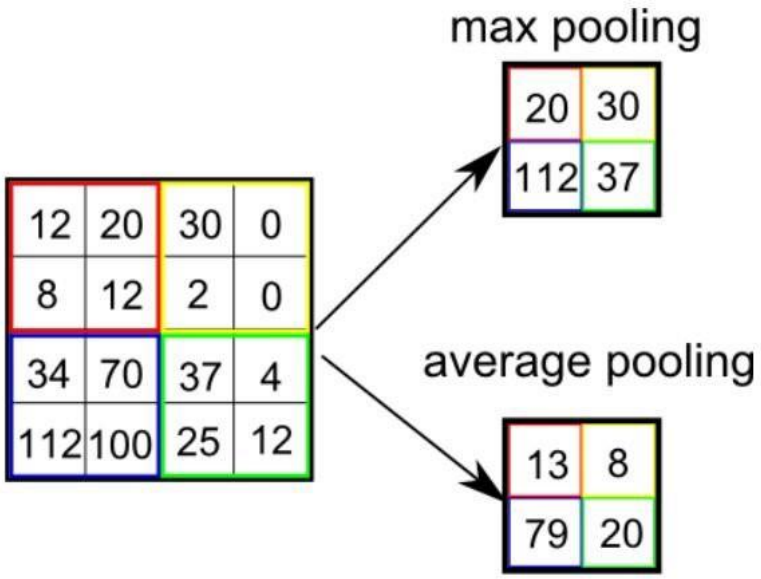
Với mỗi bộ lọc khác nhau sẽ học được những đặc trưng khác nhau của ảnh, do đó mỗi tầng convolutional ta sẽ dùng nhiều bộ lọc (CNN1 sử dụng 32 bộ lọc) để học được nhiều đặc trưng của ảnh (ví dụ biên ngang, biên dọc…). Vì mỗi bộ lọc cho ra đầu ra là 1 ma trận nên k bộ lọc sẽ cho ra k ma trận đầu ra. Ta sẽ kết hợp (tính tổng) k ma trận đầu ra này lại thành 1 ma trận đầu ra duy nhất.




- MaxPooling2D: lớp Pooling thường được dùng giữa các lớp convolutional, để giảm kích thước dữ liệu nhưng vẫn giữ được các thuộc tính quan trọng. Kích thước dữ liệu giảm sẽ giúp giảm việc tính toán trong model.

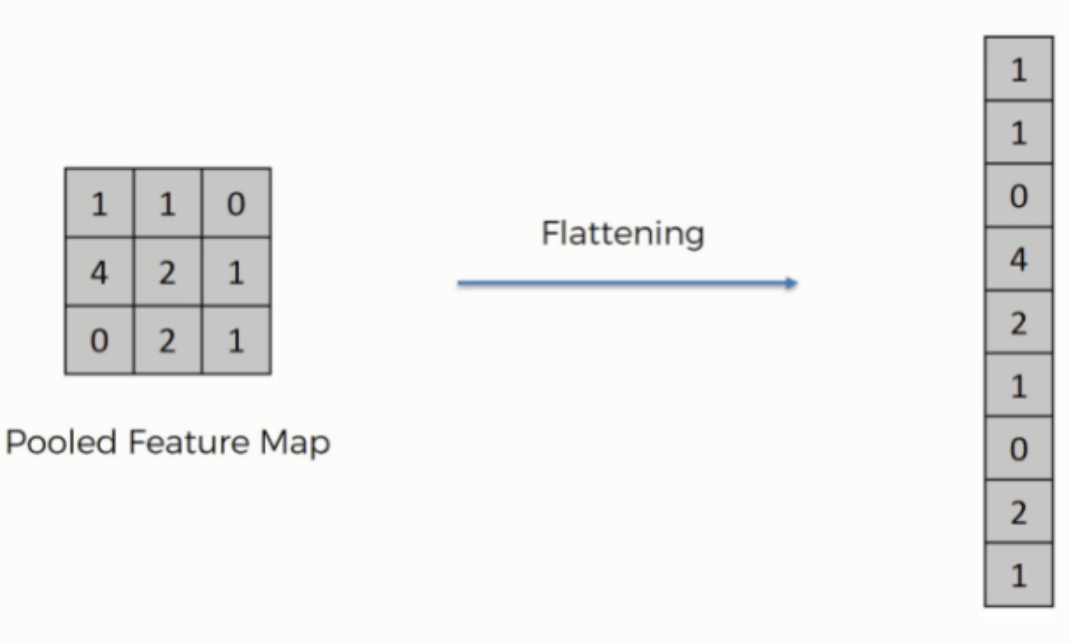
Hầu hết khi dùng pooling layer thì sẽ dùng cửa sổ trượt size=(2,2), bước dịch chuyển stride=2. Khi đó kích thước của dữ liệu sẽ giảm đi một nửa.

Có 2 loại pooling layer phổ biến là: max pooling và average pooling.




- Hàm kích hoạt Relu để loại các giá trị âm
- Flatten: chuyển ảnh từ dạng ma trận về mảng 1 chiều

Sau khi ảnh được truyền qua nhiều lớp CNN thì mô hình đã học được các đặc điểm của ảnh, khi đó output của lớp CNN cuối cùng là một ma trận, sẽ được chuyển về 1 vector một chiều.



3.5.4.8. Thiết lập các tham số để huấn luyện mô hình

```
[8]  from tensorflow.keras.optimizers import RMSprop

     # Pass learning rate using 'learning_rate' instead of 'lr'
     model.compile(optimizer=RMSprop(learning_rate=0.001),
                   loss='categorical_crossentropy',
                   metrics=['acc'])
```

- compile: chọn các tham số để huấn luyện mô hình.

- optimizer: thuật toán huấn luyện mô hình, sử dụng RMSprop hoặc 'adam' hoặc 'sgd'...

- loss: hàm tính toán sai số giữa giá trị học được và giá trị thực tế, sử dụng categorical_crossentropy trong trường hợp dự đoán nhiều lớp.

- metrics: thước đo để ta đánh giá độ chính xác accuracy của mô hình.

3.5.4.9. Huấn luyện mô hình

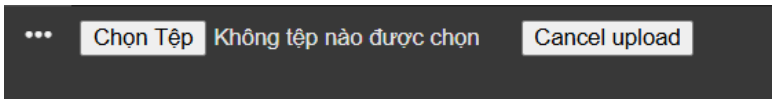
```
[9]  EPOCHS = 100
     history = model.fit(
         train_generator,
         steps_per_epoch=2,
         epochs=EPOCHS,
         verbose=1,
         validation_data=validation_generator,
         validation_steps=2
     )
```

- EPOCHS: Số vòng lặp chạy mô hình

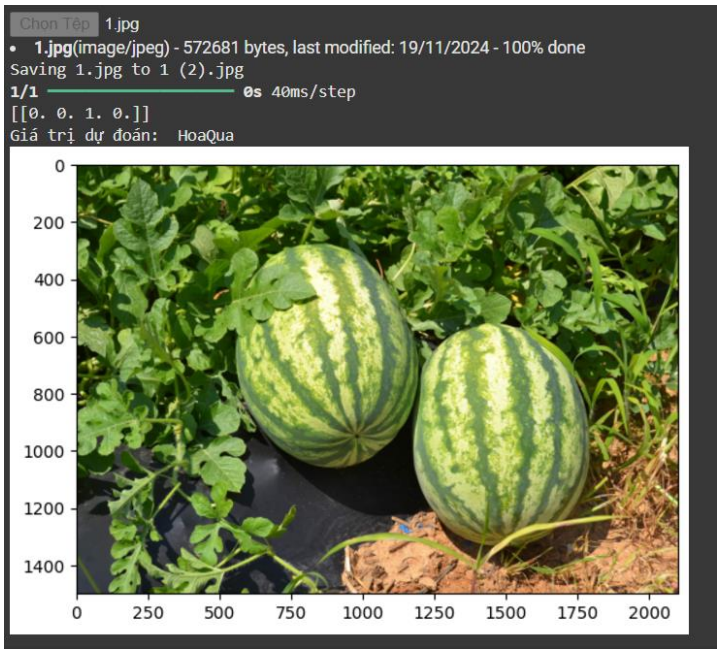
### 3.5.4.10. Sử dụng mô hình

```
[12] from google.colab import files
     from keras.preprocessing import image
     %matplotlib inline
     import matplotlib.pyplot as plt
     import matplotlib.image as mpimg

     uploaded = files.upload()

     for fn in uploaded.keys():
         # Predicting images
         path = '/content/' + fn
         # In ảnh đọc được
         plt.imshow(mpimg.imread(path))

         img = image.load_img(path, target_size=(50, 50))
         x = image.img_to_array(img)
         x = np.expand_dims(x, axis=0)
         images = np.vstack([x])
         y_predict = model.predict(images, batch_size=10)
         print(y_predict)

         print('Giá trị dự đoán: ', label[np.argmax(y_predict)])
```

- Chạy đoạn code, xuất hiện kết quả yêu cầu chọn hình ảnh để phân loại, kích chọn nút Chọn tệp, xuất hiện hộp thoại Open, chọn một ảnh để phân loại.




- Giả sử chọn ảnh 1.jpg, ta có kết quả sau:



# PHẦN IV: KẾT LUẬN VÀ BÀI HỌC KINH NGHIỆM

Đề tài đã thực hiện nghiên cứu tính toán hiệu năng cao (HPC) trên các nền tảng đám mây hàng đầu như Google, Amazon, Microsoft và các đám mây trong nước, nhằm đánh giá tiềm năng và khả năng ứng dụng của chúng. Trong quá trình nghiên cứu, nhóm đã triển khai sâu vào Google COLAB với các tài nguyên GPU và TPU, từ đó xây dựng và huấn luyện thành công một mô hình mạng nơ-ron học sâu hình có khả năng nhận diện mặt hàng thực phẩm và đồ uống. Kết quả thu được đã cho thấy tính hiệu quả và sự linh hoạt khi sử dụng đám mây cho HPC trong việc giải quyết các bài toán phức tạp như huấn luyện mô hình học máy và học sâu. Điều này mở ra tiềm năng ứng dụng mạnh mẽ trong các lĩnh vực công nghệ và khoa học dữ liệu trong tương lai.

# TÀI LIỆU THAM KHẢO


[1] BKAII, "Máy tính hiệu năng cao (High Performance Computing – HPC) là gì?" [Online]. Available: https://bkaii.com.vn/tin-tuc/987-may-tinh-hieu-nang-cao-high-performance-computing-hpc-la-gi. [Accessed: Dec. 17, 2024].

[2] Intel, "High-Performance Computing (HPC) Architecture," Intel Vietnam. [Online]. Available: https://www.intel.vn/content/www/vn/vi/high-performance-computing/hpc-architecture.html. [Accessed: Dec. 17, 2024].

[3] Amazon Web Services, "High Performance Computing on AWS," AWS. [Online]. Available: https://aws.amazon.com/vi/hpc/. [Accessed: Dec. 17, 2024].

[4] Hewlett Packard Enterprise, "What is HPC Cloud?" [Online]. Available: https://www.hpe.com/emea_europe/en/what-is/hpc-cloud.html. [Accessed: Dec. 17, 2024].

[5] Intel, "High-Performance Computing in the Cloud," Intel Corporation. [Online]. Available: https://www.intel.com/content/www/us/en/high-performance-computing/cloud.html. [Accessed: Dec. 17, 2024].

[6] HPC Wire, "Google Cloud's HPC Innovations Powering the Future of High-Performance Computing," Oct. 2, 2023. [Online]. Available: https://www.hpcwire.com/2023/10/02/google-clouds-hpc-innovations-powering-the-future-of-high-performance-computing/. [Accessed: Dec. 17, 2024].

[7] EITCA Academy, "What Are the Key Components Involved in Building an HPC Environment on Google Cloud?" [Online]. Available: https://eitca.org/cloud-computing/eitc-cl-gcp-google-cloud-platform/gcp-basic-concepts/high-performance-computing/examination-review-high-performance-computing/what-are-the-key-components-involved-in-building-an-hpc-environment-on-google-cloud/. [Accessed: Dec. 17, 2024].

[8] Amazon Web Services, "High-Performance Computing on AWS: Architecture Diagrams," [Online]. Available: https://docs.aws.amazon.com/architecture-

diagrams/latest/high-performance-computing-on-aws/high-performance-computing-on-aws.html. [Accessed: Dec. 17, 2024].