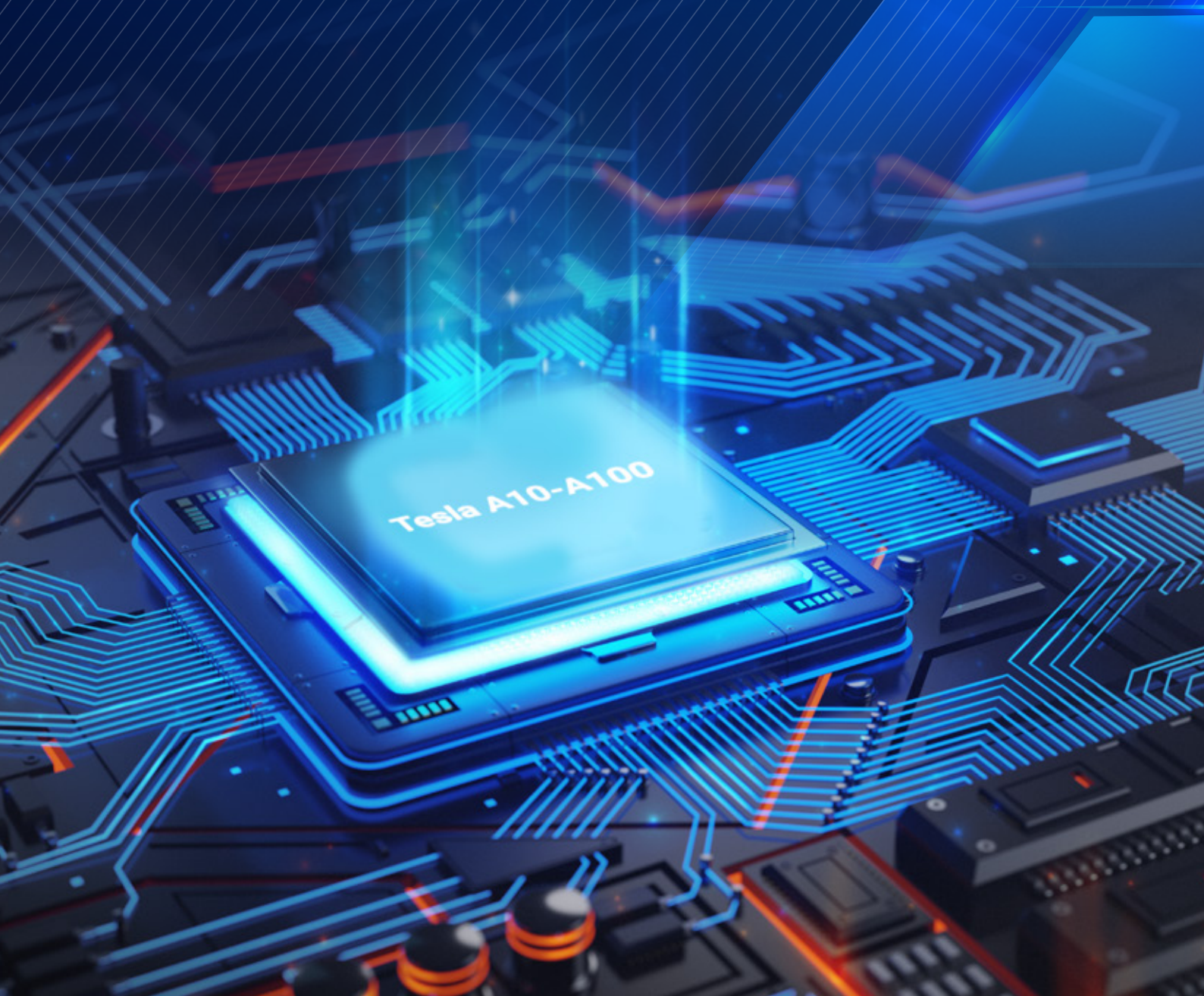
FPT CLOUD
by FPT Smart Cloud

ĐỘT PHÁ HIỆU SUẤT PHÁT TRIỂN AI
VỚI DỊCH VỤ GPU



MỤC LỤC

01

CÔNG NGHỆ AI TRÊN HẠ TẦNG TÍCH HỢP GPU

Trong lĩnh vực trí tuệ nhân tạo (AI), học máy (machine learning - ML) và học sâu (deep learning - DL) là các công cụ hiệu quả hỗ trợ các vấn đề về thuật toán phức tạp, từ phân loại đối tượng, hiểu ngôn ngữ đến hệ thống hỗ trợ tự động. Hiện nay, lĩnh vực trí tuệ nhân tạo (AI) đang được ứng dụng rất nhiều trong thực tế, từ người dùng cho đến các doanh nghiệp và tập đoàn. Với các doanh nghiệp, ML và DL được sử dụng cho nhiều ứng dụng như dự báo nhu cầu sản phẩm, đánh giá rủi ro tài chính, phát hiện lỗi sản xuất, ngăn ngừa tổn thất bán lẻ,...

Điều kiện để triển khai một dự án AI sẽ bao gồm một quy trình tính toán chuyên sâu và phức tạp, do đó việc sở hữu hạ tầng phù hợp nhằm tối ưu hóa cho các nhu cầu chuyên biệt. Điểm quan trọng nhất cho các hạ tầng có khả năng triển khai và phát triển một dự án AI chính là các máy chủ tăng tốc GPU (GPU Accelerated chuyên dụng cho phát triển AI). Việc tích hợp máy chủ GPU accelerated vào hạ tầng dùng để phát triển dự án AI cũng sẽ khiến cho quy trình làm việc của tổ chức có thể trở nên phức tạp do quy trình cung ứng quản lý và giám sát chuyên biệt. Doanh nghiệp sẽ gặp phải rất nhiều khó khăn trong quá trình tích hợp hạ tầng công nghệ AI/ML vào hệ thống máy chủ được hỗ trợ bởi GPU chuyên biệt. Trong tài liệu này, doanh nghiệp có thể biết thêm về các thách thức dễ gặp trong quá trình triển khai dự án AI trên GPU. Thách thức doanh nghiệp có thể gặp phải bao gồm:

1. Triển khai cơ sở hạ tầng đặc thù cho GPU
2. Tích hợp ứng dụng AI vào các quy trình triển khai hiện có
3. Giám sát tinh chỉnh để tối ưu hiệu quả hạ tầng GPU.
4. Hiệu quả mở rộng của các triển khai AI trên GPU.

Bộ Tài liệu là sự kết hợp giữa đội ngũ chuyên gia về phát triển ứng dụng và dự án AI của **FPT Cloud** với **NVIDIA** để đưa cho doanh nghiệp cái nhìn tổng quan về các thách thức mà doanh nghiệp sẽ dễ gặp phải trong quá trình bắt đầu triển khai các cơ sở hạ tầng chuyên biệt dành cho các dự án phát triển AI trong tương lai

04 THÁCH THỨC CHÍNH KHI TRIỂN KHAI AI TRÊN GPU

THÁCH THỨC #1: TRIỂN KHAI CƠ SỞ HẠ TẦNG ĐẶC THÙ CHO GPU

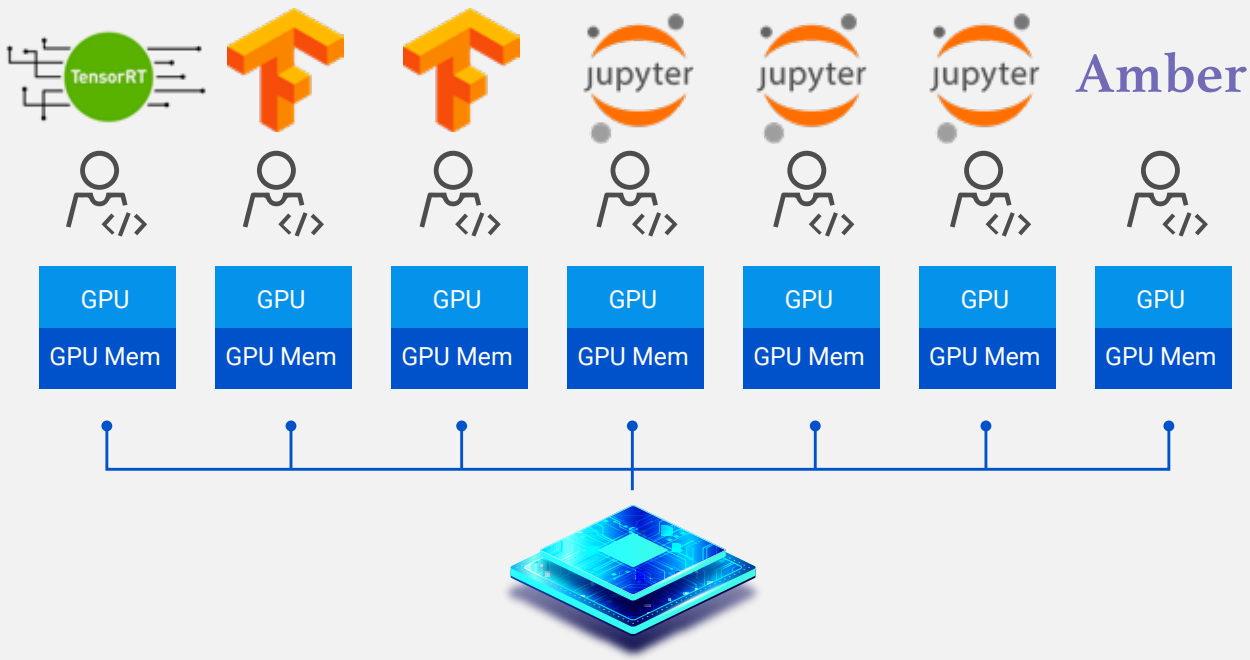
Chúng ta đều nắm rõ về các quy trình hỗ trợ cho việc cung ứng, nâng cấp và bảo trì hạ tầng máy chủ chỉ sử dụng CPU truyền thống. Ngược lại, các máy chủ GPU accelerated chỉ chuyên dụng cho việc thực hiện các thuật toán phức tạp và phát triển AI. Do đó, doanh nghiệp thường gặp khó khăn vì tính chất phức tạp trong quá trình mua sắm phần cứng, nâng cấp hay bảo trì hệ. Để đảm bảo được việc tối ưu quy trình GPU, doanh nghiệp có thể tận dụng một số giải pháp sau:

Hệ thống trao tay (turnkey System) giúp giảm bớt độ trễ khi thiết lập hạ tầng

Các giải pháp tích hợp hạ tầng hoặc các kiến trúc hạ tầng đã có sẵn mẫu thiết kế chi tiết cùng với các bài toán đánh giá cho toàn bộ hệ thống giúp loại bỏ bớt đi sự phức tạp trong quá trình triển khai hạ tầng hệ thống GPU mới của doanh nghiệp. Đối với Doanh nghiệp đang có kế hoạch để triển khai dự án phát triển hệ thống AI riêng thì việc có một kiến trúc mẫu sẽ mang đến khả năng tính toán các thuật toán, kết nối và lưu trữ trong cùng một thiết kế cơ sở hạ tầng phát triển AI được hỗ trợ bởi GPU. Các giải pháp này có sẵn dưới dạng một giải pháp cho cơ sở hạ tầng trao tay từ các nhà cung cấp đã có kinh nghiệm triển khai và xây dựng hạ tầng hỗ trợ cho AI. Các nhà cung cấp có khả năng tư vấn tùy chỉnh thiết kế cho phù hợp với nhu cầu kinh doanh cụ thể của mỗi doanh nghiệp. đồng thời, Các nhà cung cấp này cũng sẽ có các dịch vụ tư vấn thiết kế, cài đặt, vận hành và bảo trì.

Các mẫu kiến trúc hạ tầng sẽ rút ngắn thời gian để hoàn thiện một máy chủ GPU accelerated với đầy đủ chức năng. Các doanh nghiệp cũng có thể tự thiết kế và hoàn thiện hạ tầng hoặc thuê đối tác khác cung cấp một hệ thống hoàn thiện để tiết kiệm thời gian triển khai ban đầu

Tinh chỉnh linh hoạt với Multi-Instance GPU (MIG)



Multi-Instance GPU (MIG) là một tính năng mang tính cách mạng trong thiết kế kiến trúc Ampere (Ampere Architecture) mới từ NVIDIA. Sau khi GPU NVIDIA A100 được cài đặt vào hệ thống, Bộ phận quản trị hạ tầng có thể tối ưu việc sử dụng GPU và mở rộng quyền truy cập GPU cho nhiều người dùng hơn với tính năng này. Các lợi ích chính của MIG bao gồm:

1. Người quản trị hệ thống có thể chia một GPU A100 thành tối đa 7 GPU Instances khác nhau; mỗi máy GPU Instance có bộ tính toán chuyên biệt (Streaming Multiprocessors), bộ nhớ, bộ nhớ cache L2 và băng thông bộ nhớ cho QoS và cách ly phần cứng.

2. Mỗi GPU instance có khả năng chạy song song và do đó có thể thực hiện đồng thời khối lượng công việc với thông lượng và độ trễ có thể dự đoán được.

3. Với MIG, người quản trị hệ thống có thể cung cấp các GPU instance có kích thước phù hợp dựa trên yêu cầu khối lượng công việc.

Từ góc độ vận hành, các MIG instance xuất hiện dưới dạng tài nguyên GPU bổ sung trong các công cụ điều phối container như Kubernetes thông qua việc tích hợp thiết bị NVIDIA với Kubernetes. Quản trị viên có thể sử dụng MIG như một công cụ tinh chỉnh năng suất để nhanh chóng mở rộng quyền truy cập GPU cho nhiều người dùng và duy trì mức sử dụng GPU cao trên trung tâm dữ liệu của họ.

THÁCH THỨC #2:
TÍCH HỢP ỨNG DỤNG AI VÀO CÁC QUY TRÌNH TRIỂN KHAI HIỆN TẠI

Khi triển khai ứng dụng, doanh nghiệp thường làm theo hai cách thức sau:

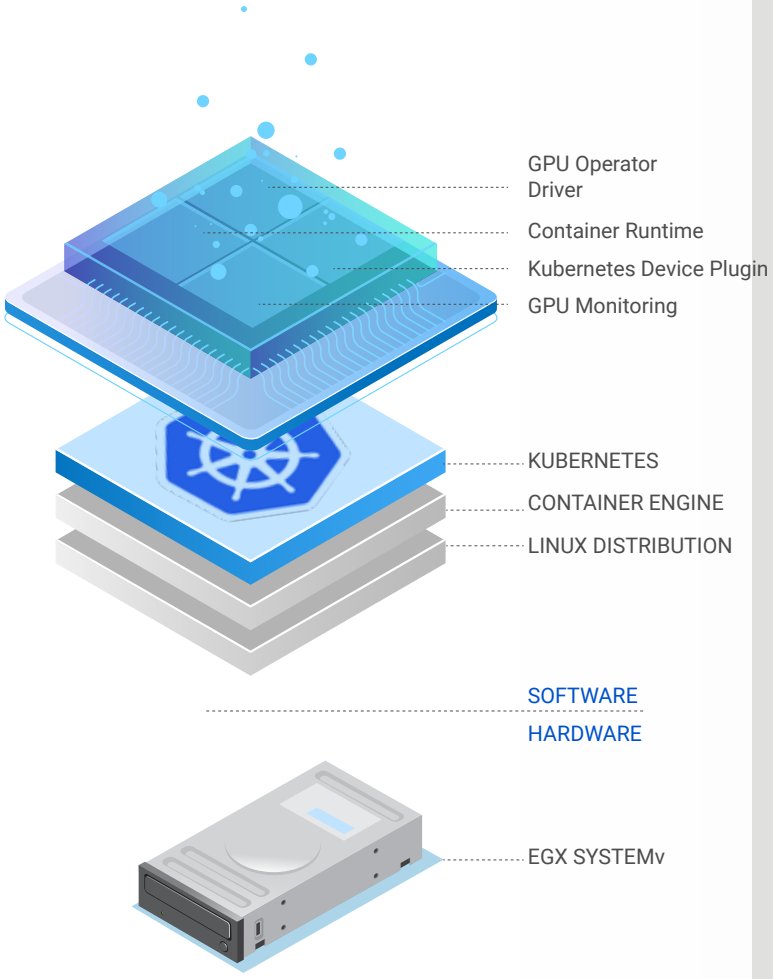
1. Máy chủ vật lý: Các ứng dụng được chứa trong containers và chạy trực tiếp trên các máy chủ.

2. Máy chủ ảo: Các ứng dụng được cài đặt trên máy ảo và chạy trên các phần mềm giám sát máy ảo trên máy chủ vật lý.

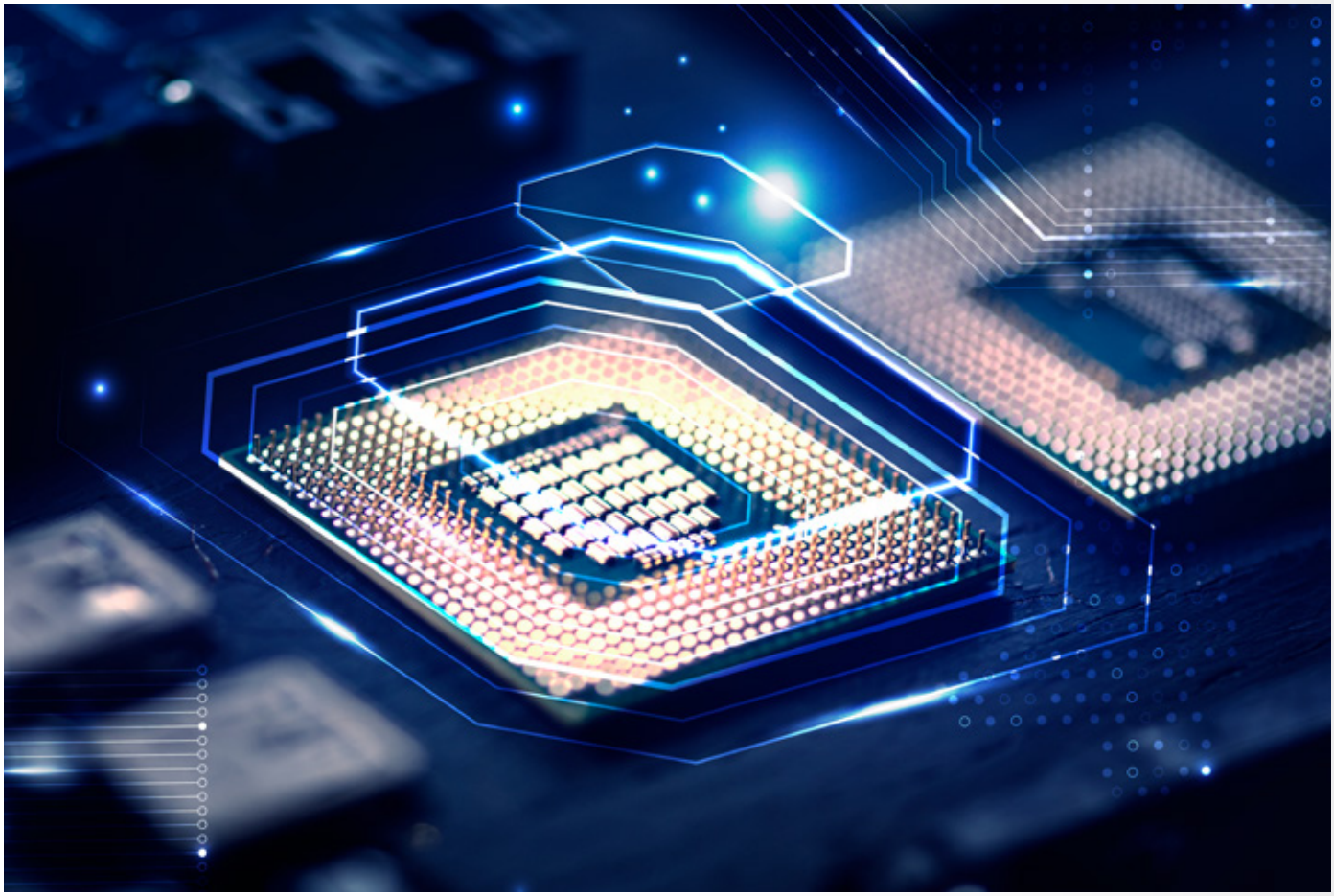
Trước tiên, hãy cùng xem xét các container trước. Trong lĩnh vực này, Kubernetes đang dần trở thành nền tảng điều phối container được nhiều doanh nghiệp tin dùng. Container và Kubernetes cho phép DevOps triển khai các ứng dụng trên trung tâm dữ liệu, Public Cloud, Hybrid Cloud, và Edge

Thiết lập Kubernetes Cluster nhằm hỗ trợ hạ tầng GPU

Các GPU trên máy chủ phải được cấu hình với driver, container runtime và các thư viện khác trước khi chúng có thể được sử dụng làm tài nguyên bởi Kubernetes. Việc cấu hình và quản lý chúng thủ công rất tốn thời gian, thậm chí có thể dẫn đến sai sót.

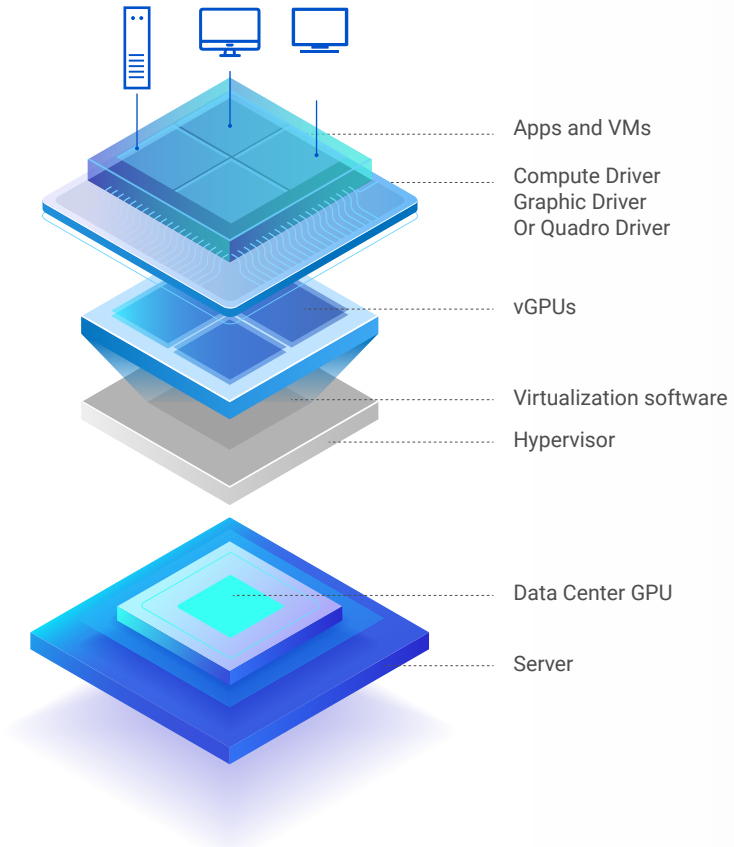


Doanh nghiệp sẽ cần một dịch vụ giống như là một Kubernetes trừu tượng (Kubernetes abstraction) giúp tự động hóa việc cấu hình GPU node bằng cách cài đặt thiết bị GPU K8s và các công cụ giám sát GPU trên toàn bộ cluster. Ví dụ như, GPU Operator sẽ giúp đơn giản hóa cả việc triển khai ban đầu và việc quản lý các thành phần hệ thống hạ tầng bằng cách tập trung lại và sử dụng Kubernetes tiêu chuẩn. Các nhà quản trị hạ tầng hay phát triển hệ thống có thể sử dụng các API Kubernetes tiêu chuẩn để tự động hóa và quản lý các thành phần này, bao gồm cả lập phiên bản và nâng cấp. GPU operator cũng cho phép người quản trị hệ thống và DevOps tách cấu hình GPU khỏi việc cung ứng máy chủ. Điều đó cho phép quản trị viên chỉ phải dùng một golden OS image duy nhất cho cả máy chủ CPU và GPU, làm giảm chi phí hoạt động một cách hiệu quả.



Ảo hóa máy chủ cho GPU

Ảo hóa máy chủ được hầu hết doanh nghiệp sử dụng nhờ vào tính linh hoạt cho ứng dụng và tối ưu hạ tầng. Đối với doanh nghiệp, việc triển khai các ứng dụng trong máy ảo trở thành thông lệ cơ bản. Nhằm duy trì sự ổn định, giảm thiểu việc triển khai một lần, bộ phận Vận hành có thể triển khai các ứng dụng AI được gia tốc bởi GPU trên máy ảo.



NVIDIA vComputeServer là công nghệ ảo hóa GPU cho phép ảo hóa máy chủ truyền thống trên với các GPU của NVIDIA. Phần mềm được tích hợp với nhiều nền tảng hypervisor phổ biến như VMware vSphere, Red Hat Virtualization, KVM và Nutanix. vComputeServer cho phép các ứng dụng AI có thể tối ưu sử dụng GPU, bao gồm các container trong máy ảo trên GPU, dù là trên toàn bộ GPU hay trong các MIG instance. Nó cung cấp hiệu suất GPU với các lợi ích về bảo mật hypervisor, quản lý, giám sát và hoạt động, bao gồm cả chuyển dịch trực tiếp.

THÁCH THỨC #3:
GIÁM SÁT TINH CHỈNH ĐỂ TỐI ƯU HIỆU QUẢ HẠ TẦNG GPU

Bộ phận Vận hành liên tục theo dõi hiệu suất của ứng dụng và cơ sở hạ tầng để đáp ứng mục tiêu SLA. Với các ứng dụng sử dụng mô hình ML và DL, các số liệu chính sau đây giúp đánh giá hiệu suất của mô hình và hạ tầng GPU.

1. Độ trễ:
Là thời gian cần thiết để mô hình xử lý dữ liệu đưa vào và trả về thông tin dự đoán. Đối với các ứng dụng chạy theo thời gian thực, độ trễ phải thấp nhất có thể.

2. Thông lượng:
Đây là số lượng yêu cầu suy luận có thể được xử lý trong một đơn vị thời gian. GPU là bộ xử lý song song khổng lồ và có thể đạt được thông lượng cao mà không ảnh hưởng đến độ trễ.

GPU là công cụ tăng tốc chuyên dụng cho các dự án AI hoặc các công việc đòi hỏi tốc độ phân tích và xử lý nhanh, do đó các nhà quản lý vận hành thường tìm cách để khai thác tối đa giá trị từ tài nguyên mang đến hiệu suất cao này. Một thước đo truyền thống được dùng để đánh giá là việc sử dụng (cả mức đỉnh và mức trung bình) để đảm bảo rằng tài nguyên mang lại giá trị thực tế cho trung tâm dữ liệu.

Bảng dưới đây mô tả các số liệu chính để theo dõi khi cải thiện và tối ưu hóa hiệu suất ứng dụng trên GPU.

Tối ưu GPU (Quy mô mỗi GPU và cluster)

Tỷ lệ phần trăm thời gian GPU được sử dụng trong một khoảng thời gian. Giúp lên kế hoạch và phân bổ GPU để đáp ứng nhu cầu ứng dụng.

Sử dụng TensorCore

TensorCore trong GPU NVIDIA cung cấp hiệu suất cao hơn với đầy đủ các tính năng chính xác. Việc sử dụng TensorCore cao mang lại hiệu quả cao hơn trong huấn luyện và suy luận có hiệu suất cao.

Sử dụng bộ nhớ và băng thông

Việc sử dụng bộ nhớ giúp lên kế hoạch dung lượng và phân bổ GPU có kích thước phù hợp. Sử dụng băng thông có thể được dùng để tối ưu hóa hiệu suất ứng dụng.

Sử dụng NVLink/PCIe

Sử dụng dữ liệu qua giao diện NVLink và PCIe của GPU có thể được sử dụng để tối ưu hóa hiệu suất ứng dụng

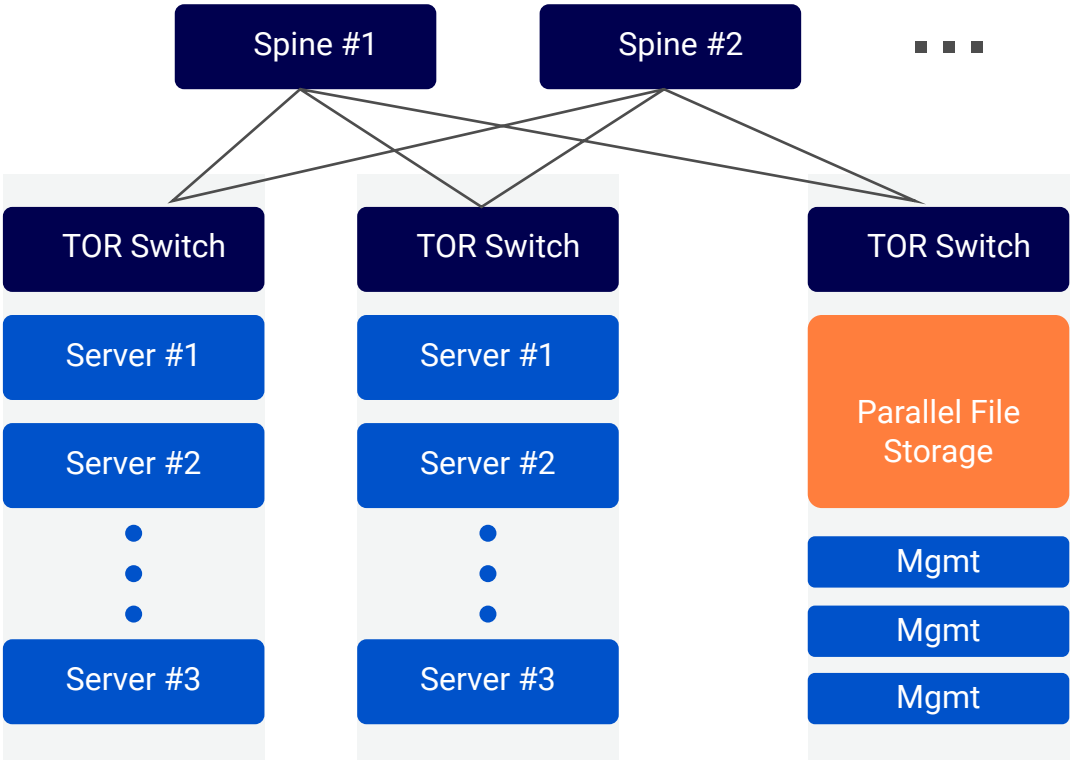
Mức tiêu thụ điện năng, nhiệt độ và lỗi GPU

Các số liệu này được sử dụng để theo dõi hiện trạng của GPU.

Bộ phận vân hành có thể dễ dàng hiển thị các chỉ số trong quá trình sử dụng GPU với các công cụ như Grafana. Nếu như doanh nghiệp đang sử dụng các máy chủ ảo có khả năng gắn thêm GPU để chạy các dự án, thì việc chạy dự án, tùy chỉnh GPU và theo dõi hiệu suất hoàn toàn có thể sử dụng trên cùng một giao diện giống như hệ thống Portal của FPT Cloud hiện tại.

THÁCH THỨC #4:
TRIỂN KHAI LINH HOẠT VÀ MỞ RỘNG CỦA DỰ ÁN AI VỚI GPU

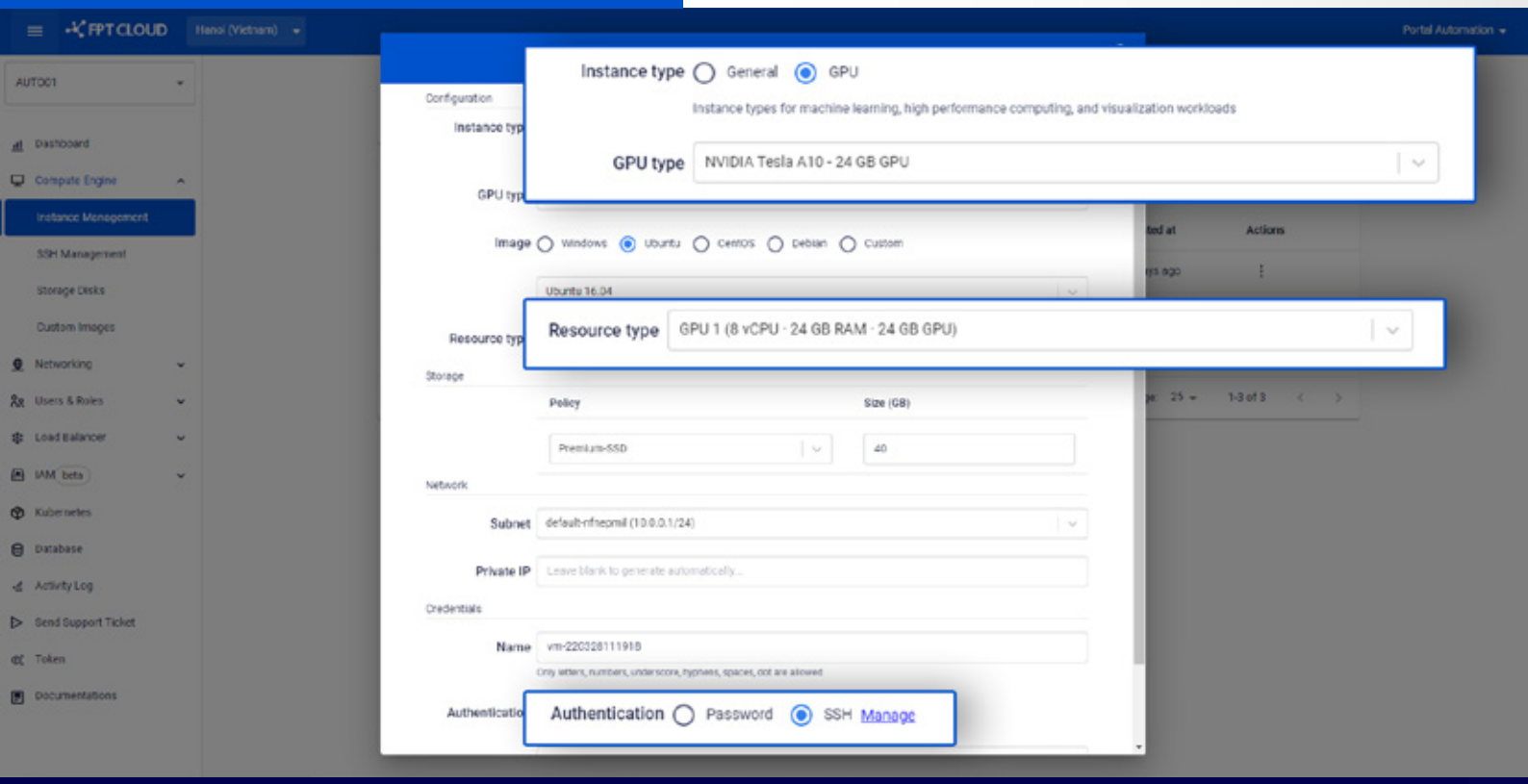
Năng lực tính toán cần thiết cho quá trình tiền xử lý dữ liệu và huấn luyện AI đang phát triển theo cấp số nhân vì khối lượng dữ liệu cũng đang phát triển ở tốc độ tương ứng. Suy luận AI (AI Inferencing) cũng đang chứng kiến sự phát triển bùng nổ với nhiều ứng dụng dụng cùng số lượng lớn người dùng cuối. Thực trạng này đòi hỏi phải nâng cấp hệ thống GPU để huấn luyện hay mở rộng quy mô GPU để suy luận AI. Tuy nhiên, việc nâng cấp/mở rộng rất phức tạp. Các giải pháp truyền thống để mở rộng cluster không phát triển theo đường thẳng và giảm hiệu suất khi các máy chủ được thêm vào hệ thống. Kiến trúc của các máy chủ, cách thức tính toán, lưu trữ và quản lý các máy chủ không được kết nối đồng nhất với nhau và với các cluster khác - ảnh hưởng đến hiệu suất mở rộng.



Để bổ sung cho phần cứng, các container trong kiến trúc microservices do Kubernetes điều phối có thể mở rộng hiệu quả các ứng dụng lên tới hàng chục nghìn GPU. Trên các đám mây công cộng, việc tự động mở rộng các GPU instance có sẵn trong FPT Cloud, AWS GKE, AKS và các ứng dụng khác.

Về dịch vụ FPT GPU Server

FPT GPU Server là dịch vụ máy chủ ảo với bộ xử lý GPU chuyên dụng giúp tăng tốc xử lý đa tác vụ tính toán cũng như đồ họa phức tạp với hàng nghìn tỷ phép toán đồng thời (30-31 TFLOPS). Dịch vụ là sự kết hợp hoàn hảo giữa tính linh hoạt, hiệu năng vượt trội của FPT Cloud Server với GPU Card mạnh mẽ nhất thị trường NVIDIA Tesla A10 và A100. Dịch vụ FPT GPU Server góp phần hoàn thiện hệ sinh thái điện toán đám mây mạnh mẽ của FPT Cloud, với đa dạng dịch vụ tiện ích từ hạ tầng (IaaS) đến nền tảng (PaaS), dịch vụ (SaaS)



Khởi tạo GPU Instance trên hệ thống Portal của FPT Cloud

Khách hàng sử dụng dịch vụ FPT GPU Server có thể lựa chọn đa dạng các mô hình dịch vụ từ mô hình GPU Instance standalone (FPT Cloud cung cấp máy chủ ảo cho khách hàng kết nối trực tiếp vào sử dụng), GPU Instance tích hợp VPC (triển khai GPU Instance ngay trong VPC của khách hàng) đến Multi GPU Instance (sử dụng GPU Instance có nhiều card GPU trên 1 máy chủ).

FPT Cloud hướng tới trở thành nền tảng chuyển đổi số, tích hợp các công nghệ mới nhất trên thế giới phù hợp với những đặc thù của môi trường kinh doanh tại Việt Nam, đưa doanh nghiệp Việt trở thành doanh nghiệp công nghệ, tăng tốc đổi mới sáng tạo, từ đó tạo ra bước nhảy vọt về năng suất lao động, trải nghiệm khách hàng, đáp ứng nhu cầu kinh doanh trong thời đại số.

Hanoi: No 10, Pham Van Bach street, Dich Vong ward, Cau Giay district
HCM: 3rd floor, PJICO building, 186 Dien Bien Phu street, district 3
info@fptcloud.com
fptcloud.com