Triton：数据挖掘

引言

Triton是一种开源的深度学习推理框架，由NVIDIA公司开发和维护。它被设计用于在图形处理单元（GPU）上进行高效的深度习模型推理。
Triton最早于2019年发布，旨在提供一个灵活且可扩的推理服务平台，以满足不同应用场下的需求。它支持多种深度学框架，包括TensorFlow、PyTorch和ONNX等，并提供了高性能的推理引擎，可以在大模部署中实现低延迟和高吞吐量。

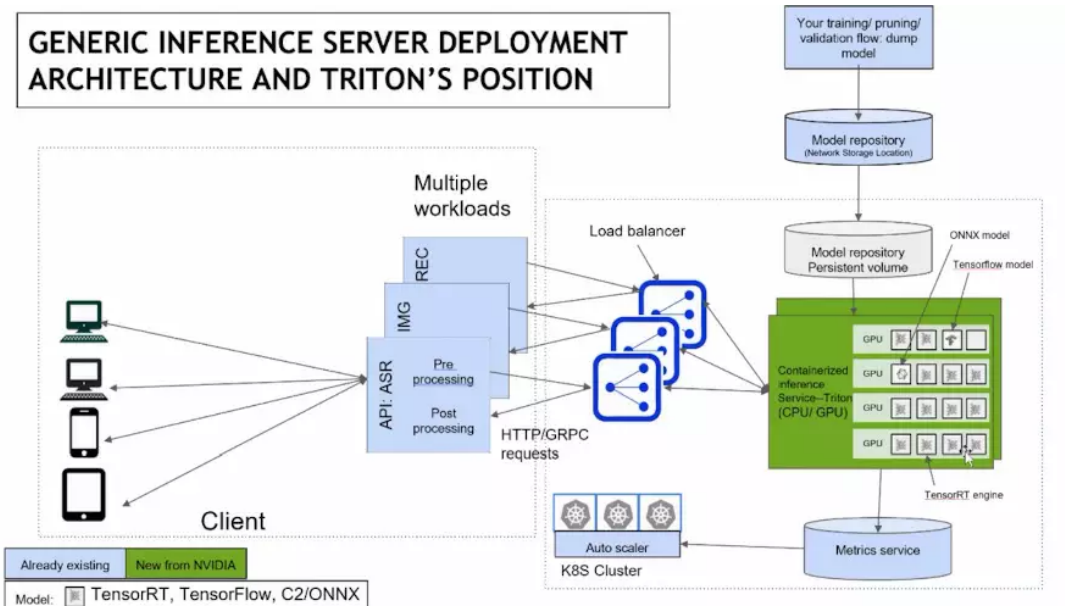
triton概述

Triton的特点和功能

Triton是一个开源的深度学习推理框架，由NVIDIA开发和维护。它主要用于在GPU上进行高能的深度学习模型推理。以下是Triton的一些特点功能：

1. 高性能推理：Triton利用GPU并行计算能力，实现了高性能深度学习模型推理。它通过优化模加载、内存管理和计算流程，供了低延迟和高吞吐量的推理性能。
2. 多模型支持Triton可以同时加载和管理多个深度习模型，并为每个模型提供独立的推理服务。这使得在同一服务器上部署和运行多个型变得更加容易和高效。
3. 灵活的部署选项：Triton支持种部署选项，括单机部署、集群部署和分布式部署。它可以根据求进行灵活的扩展和配置，以满足不同规模和复杂度的应用场景。
4. 多框架兼容性：Triton与多个主流深度学习框架兼容，包括TensorFlow、PyTorch和ONNX等。这意味着您可以使用己喜欢的框架来训练模型，并使用Triton进行推理，而无需修改现有的模型代码。
5. 动态批处理支持：Triton支持动态批处理，即可以在推理过中根据输入数据的大小自动调整批处理大小。这使得在处理不同大小的输入能够更好地利用资源，提高推理效率。
6. 强大的监控和管理功能：Triton供了丰富的监控和管理功能，包括性指标监控、模版本管理、请求日志记录等。些功能可以帮助用户更好地了解和管理推理系统的状态和性能。
7. 弹性缩：Triton支持水平和垂直扩展可以根据负载情况自动调整模的实例数量和计资源。这使得您可以根据需求灵活地扩展缩减模型的推理能力。

总之，Triton是一个功能大且灵活的深度学习理框架，它通过优化性能、提供多模型支持和灵活的部署选项帮助用户高效地部和运行深度学习模型。


GENERIC INFERENCE SERVER DEPLOYMENT ARCHITECTURE AND TRITON'S POSITION
Your training/ pruning/ validation flow: dump model
Model repository
(Network Storage Location)
Multiple workloads
REC
IMG
API: ASR
Pre processing
Post processing
Load balancer
HTTP/GRPC requests
Client
Model repository
Persistent volume
ONNX model
Tensorflow model
Containerized inference Service--Triton (CPU/ GPU)
GPU
GPU
GPU
GPU
TensorRT engine
Auto scaler
K8S Cluster
Metrics service
Already existing
New from NVIDIA
Model:
TensorRT, TensorFlow, C2/ONNX