Analytics Vidhya

The DataHour on

Optimizing Real-Time ML Inference with NVIDIA Triton Inference Server

- Sharmili Srinivasan (ML Engineer at PayPal)

Target Audience and What to expect?

- **Target Audience**
  - Students, Freshers, Beginners
- **Prerequisites**
  - Docker Fundamentals, Python Basics, ML Basics
- **What to expect?**
  - Introductions to Real Time Inferencing and Triton
  - Highlights on Triton’s distinguishing features
  - Hands-on basic setup
  - Demo on quick-win optimizations
  - Glimpse of advance optimization features

Table of contents

Brief about me..

Sharmili Srinivasan

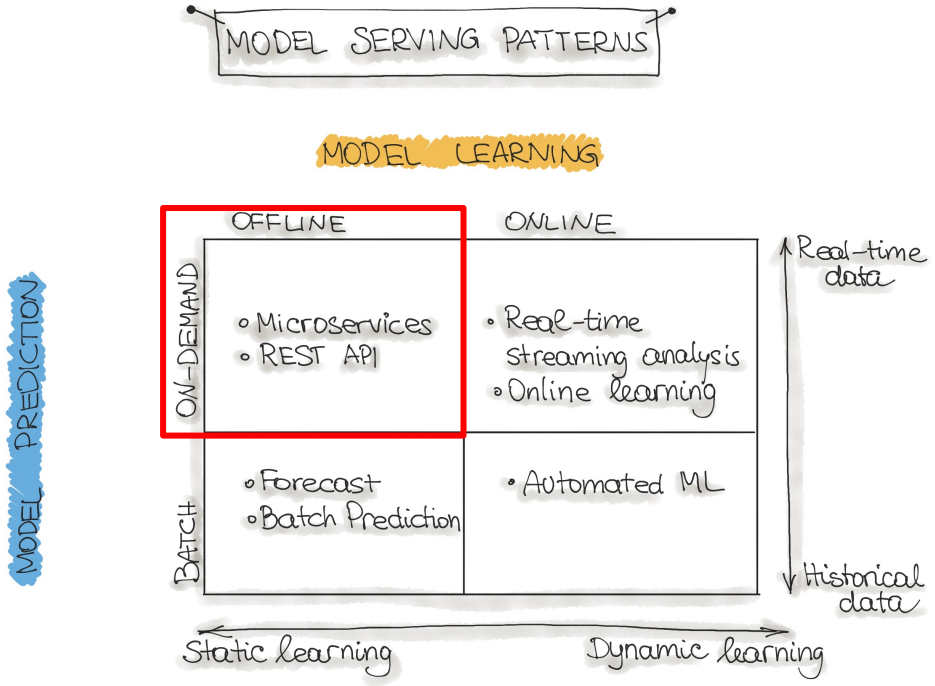
- ML Engineer @ PayPal
- 9+ years of industry experience
- Programmer by passion and profession
- Areas of interest and experience:
  - ML Engineering
  - MLOps (E2E scaled ML model deployments)
  - Big Data Engineering
- Reach me

https://www.linkedin.com/in/sharmili-srinivasan-40925a41/
sharmilisrinivasan@gmail.com
https://github.com/sharmilisrinivasan

Analytics Vidhya

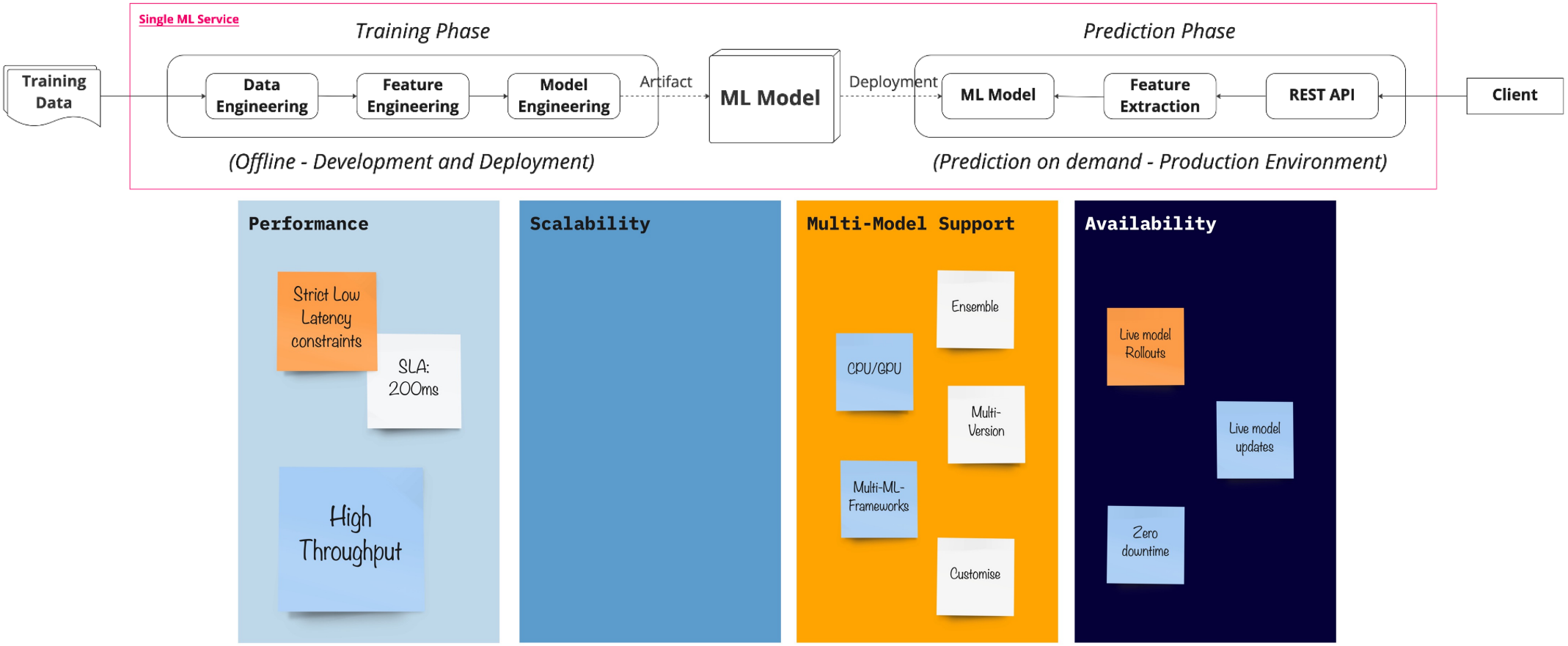
Real Time Inferencing

(Near) Real Time Inferencing –Introduction



Source: https://ml-ops.org/content/three-levels-of-ml-software#code-deployment-pipelines

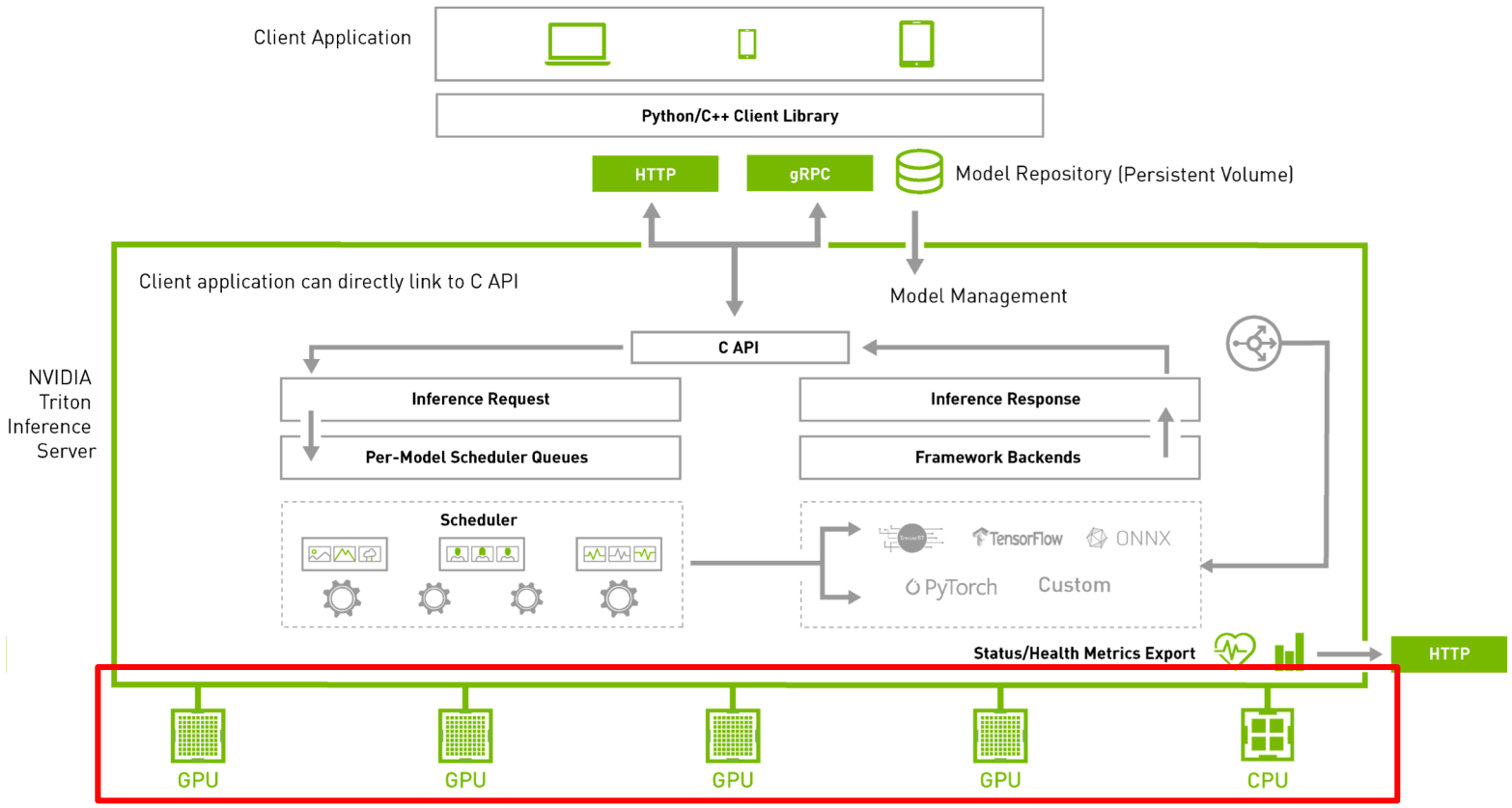
(Near) Real Time Inferencing – Expectations



Analytics Vidhya

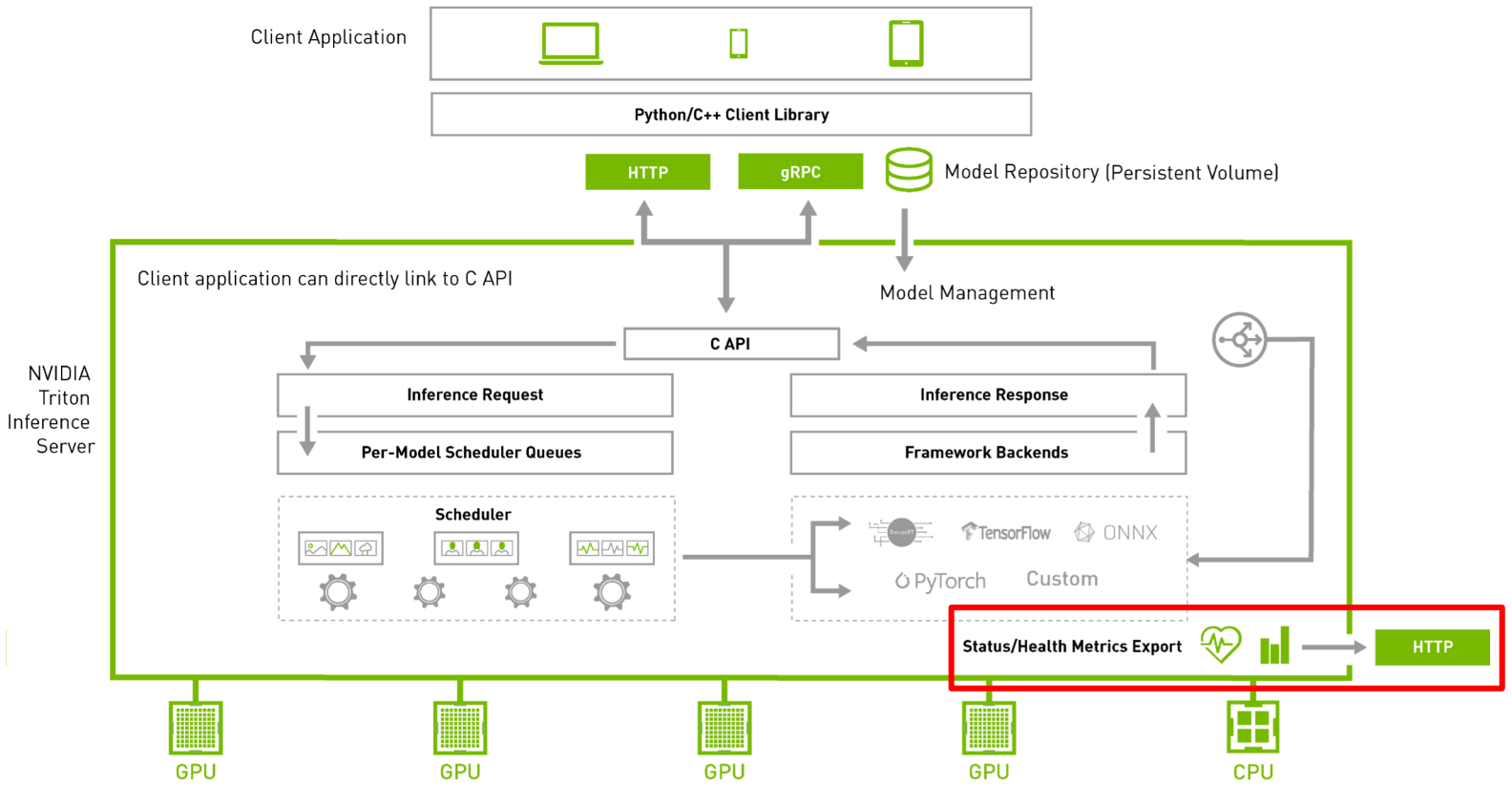
Why Triton Inference Server?

Why Triton Inference Server?



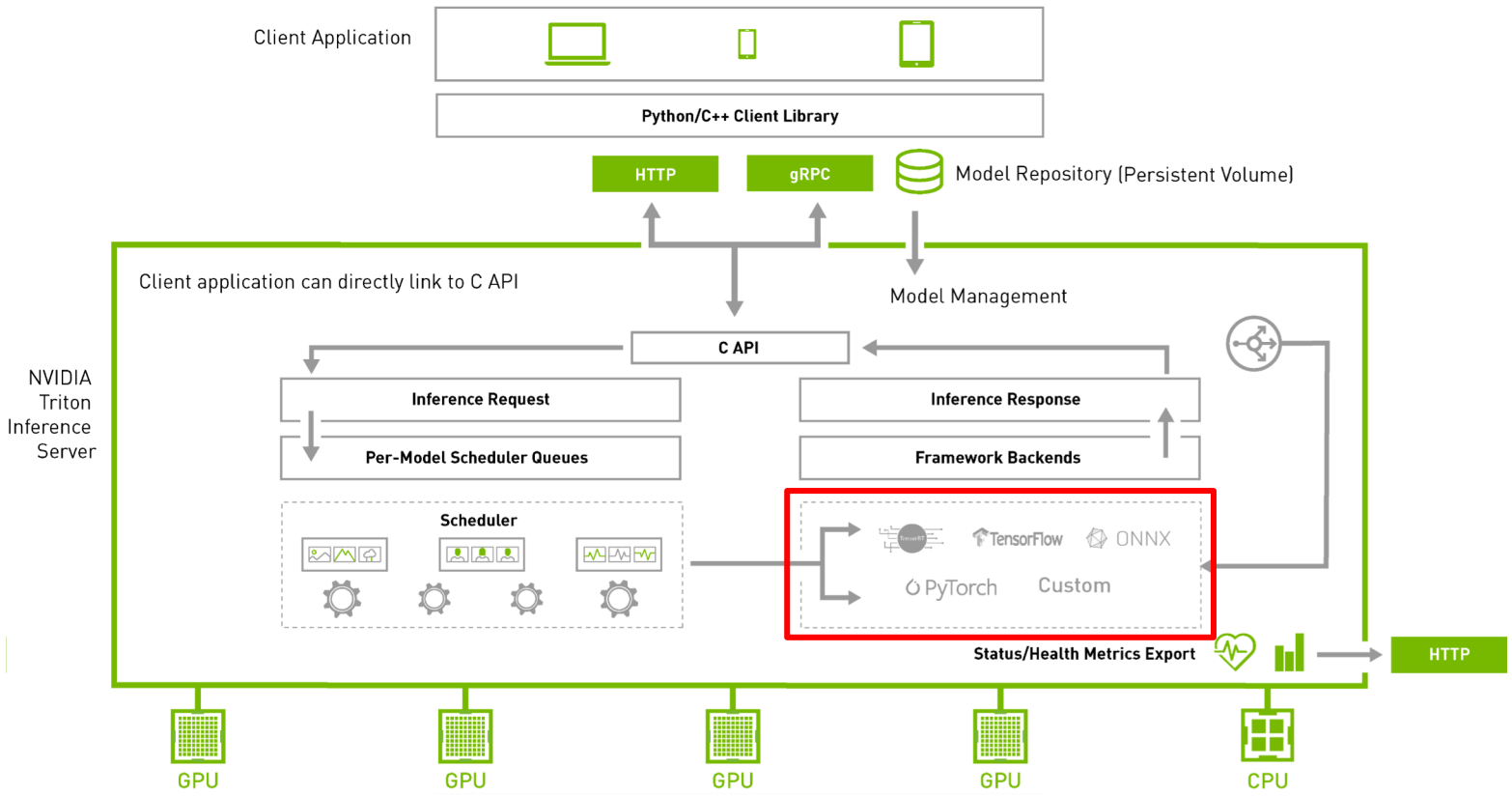
Source: https://developer.nvidia.com/nvidia-triton-inference-server

Why Triton Inference Server?



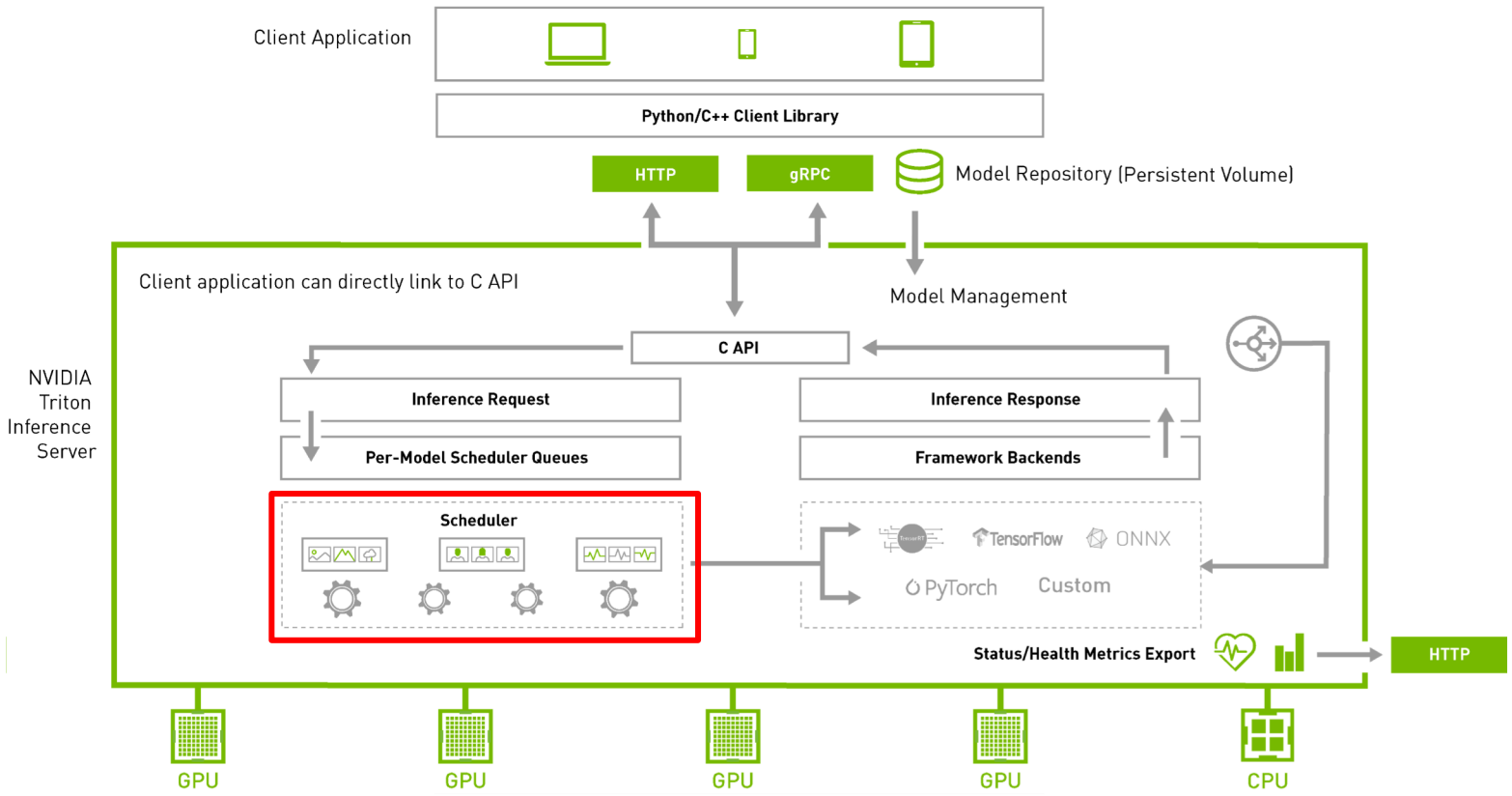
Source: https://developer.nvidia.com/nvidia-triton-inference-server

Why Triton Inference Server?



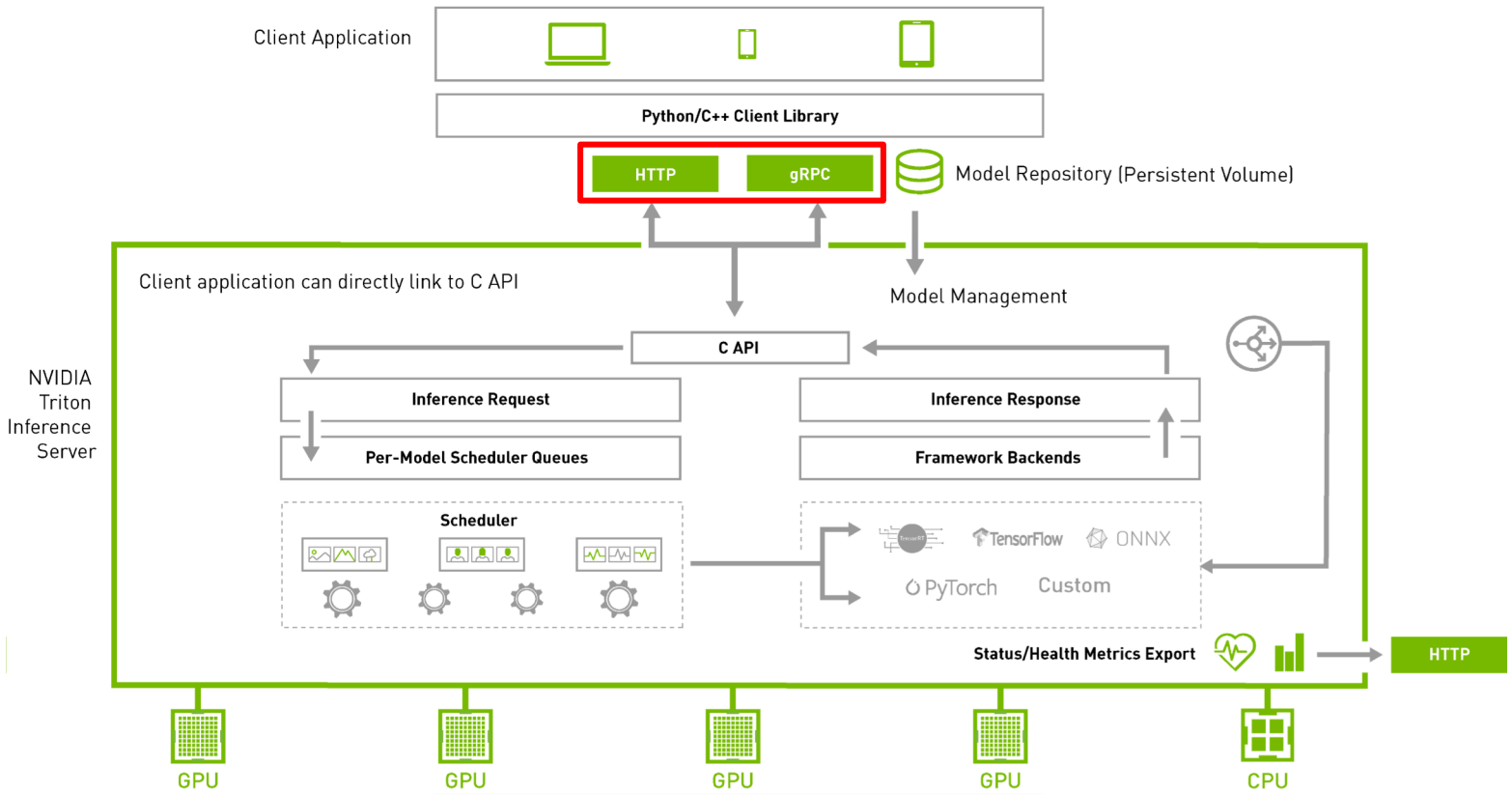
Source: https://developer.nvidia.com/nvidia-triton-inference-server

Why Triton Inference Server?



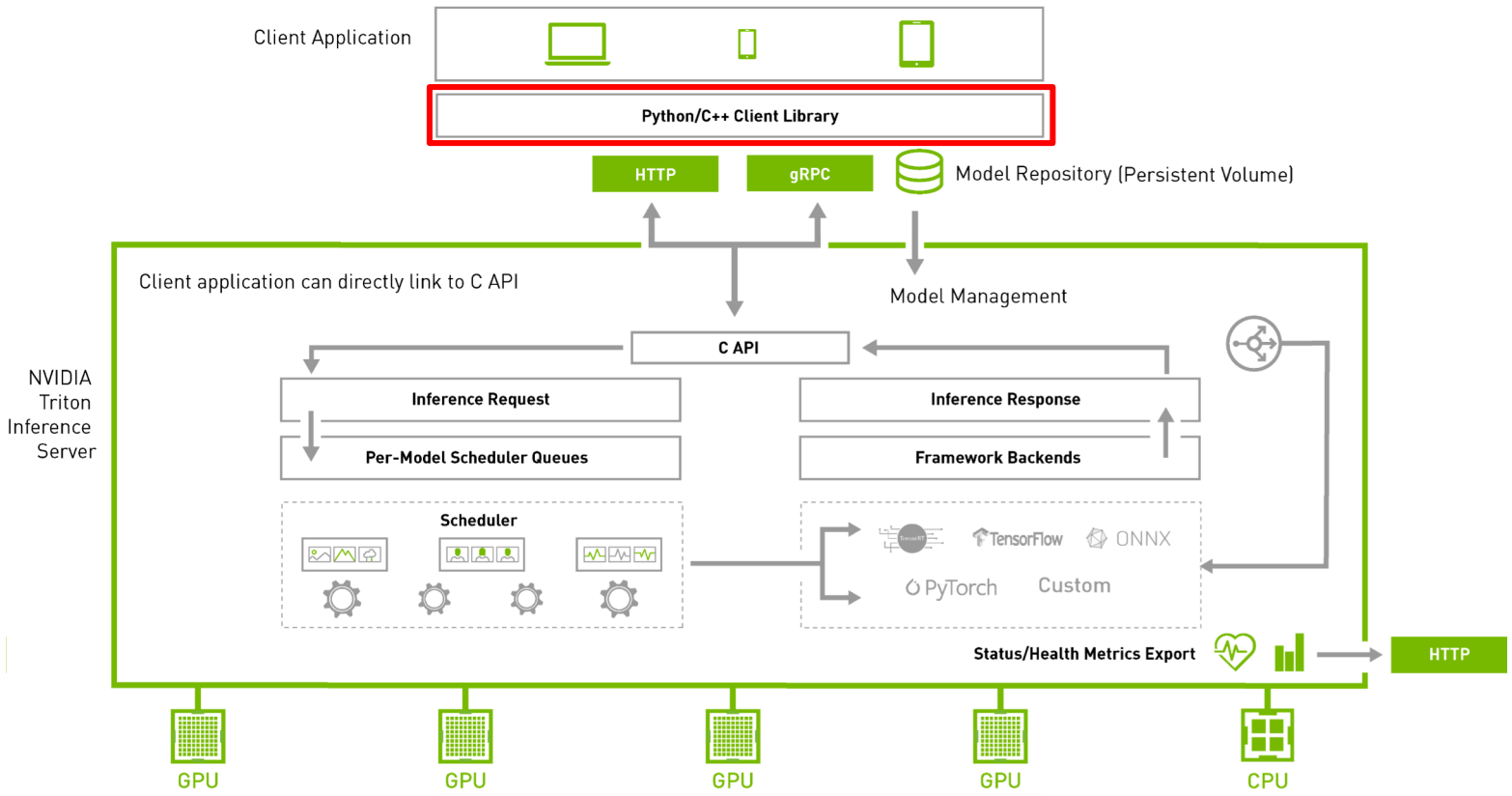
Source: https://developer.nvidia.com/nvidia-triton-inference-server

Why Triton Inference Server?



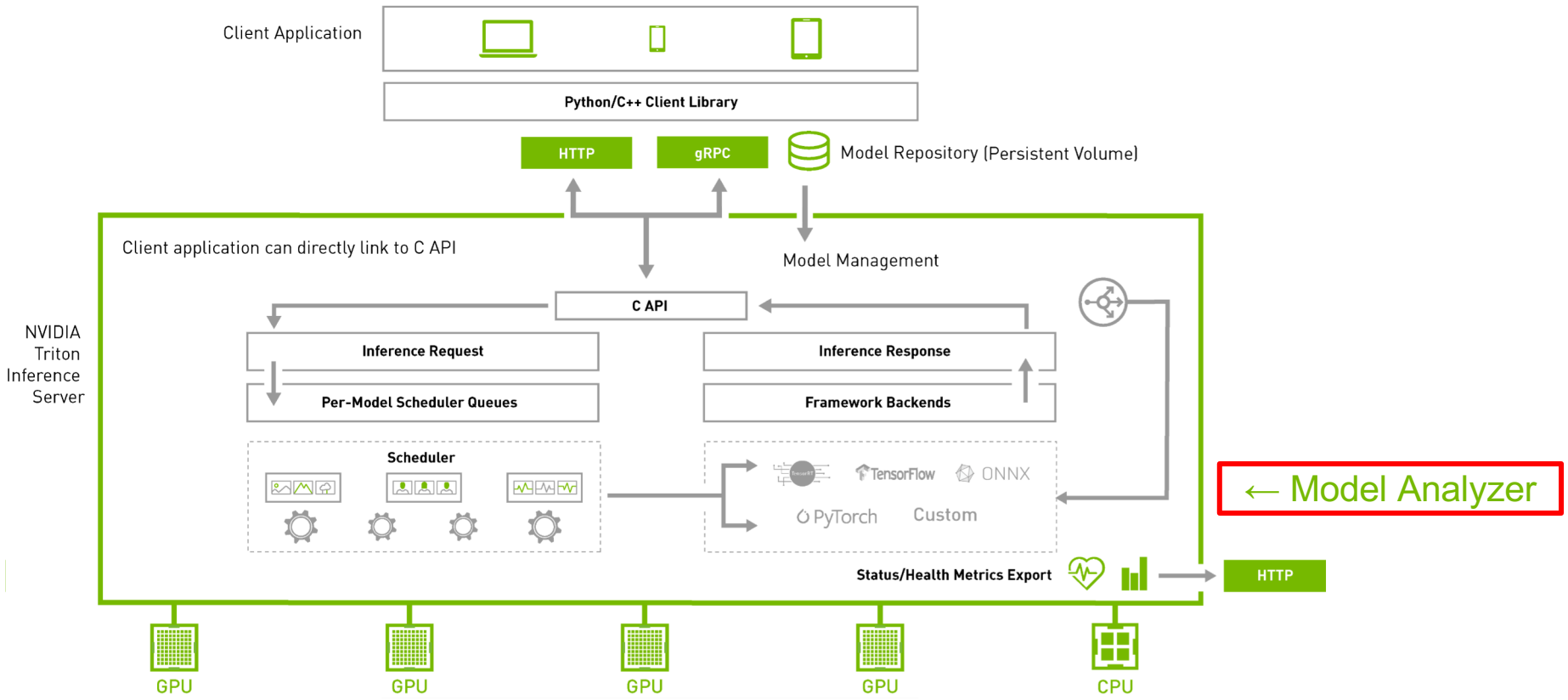
Source: https://developer.nvidia.com/nvidia-triton-inference-server

Why Triton Inference Server?



Source: https://developer.nvidia.com/nvidia-triton-inference-server

Why Triton Inference Server?

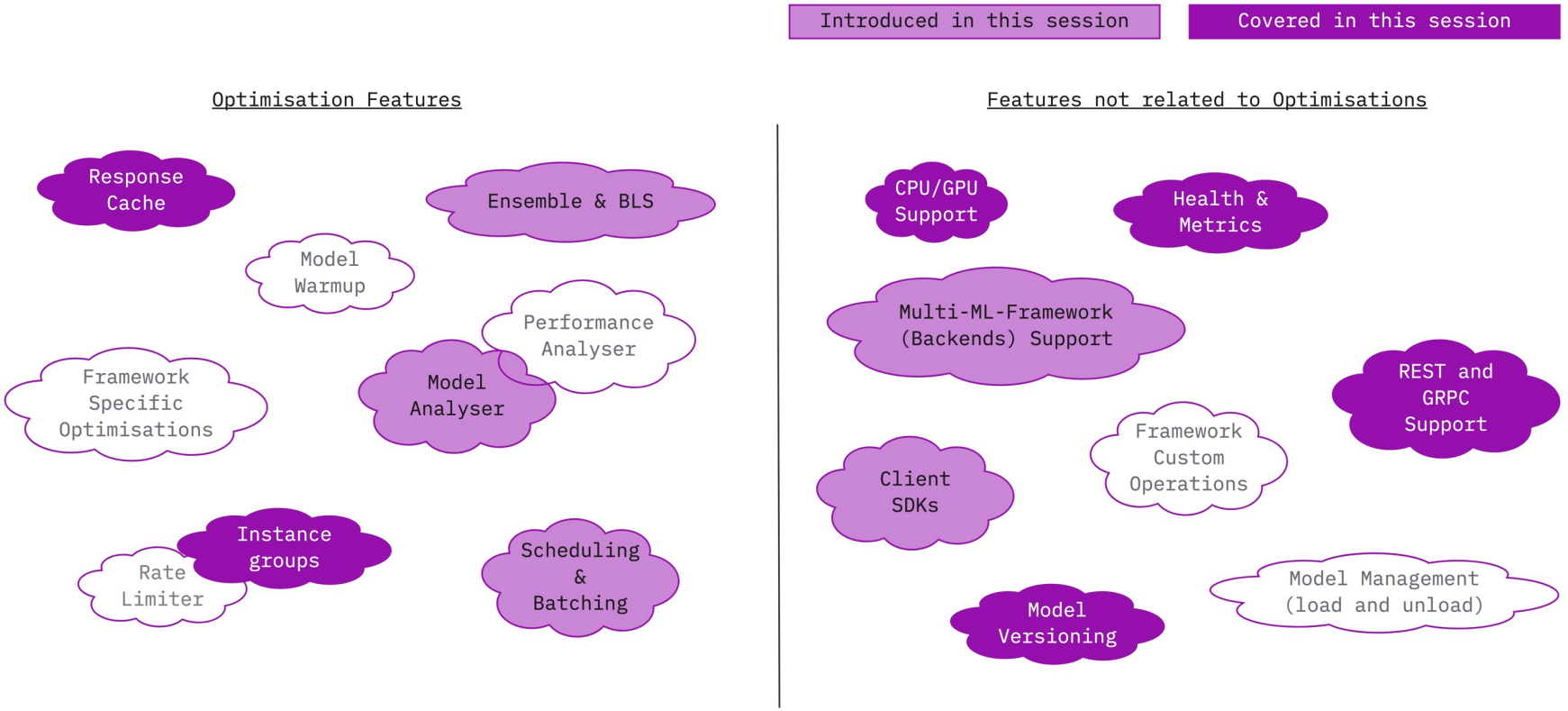


Source: https://developer.nvidia.com/nvidia-triton-inference-server

Analytics Vidhya

Triton’s Features - Highlights

Triton's Features - Highlights



Source: https://github.com/triton-inference-server/server/blob/main/docs/README.md

Analytics Vidhya

Hands-on Time

Analytics Vidhya

DataHour : Q&A

Analytics Vidhya

Thank you!