

一、Triton 总体架构

服务端：模型仓库->backend->硬件

从模型仓库加载模型，根据模型的类型，选择特定的 **backend**，将模型运行在特定的硬件上。

客户端：编程语言->query->服务端

客户端可以使用 **python** 或者 **C++**等编程语言的库，通过 **HTTP**、**gRPC** 协议，或者直接使用 **C API** 进行调用。服务端收到请求后，会调度器会调度请求给模型进行处理，返回推理结果。

二、需要了解的基本内容

- 模型仓库准备；
- 模型配置；
- Triton Server 启动；
- 配置组合模型；
- 客户端发送 Requests；

2.1 模型仓库准备

必须符合以下结构：

- 一级目录：仓库名称，如 **model_repository**。
- 二级目录：具体模型名字，如 **densenet_onnx**。
- 三级目录：
 - 版本号目录：1，表示版本 1，可以有多个版本。里面放模型文件，如 **model.onnx**；
 - **config.pbtxt**：模型配置参数，规定模型运行时的行为；
 - **label** 文件：（可选）将分类模型的输出，转为文件里指定的标签。

2.1.1 模型目录细节

版本号目录：

- 需要包含模型文件，注意格式；
- 使用模型版本需要和版本目录名保持一致。

配置文件：

为模型和服务定义一系列配置参数。

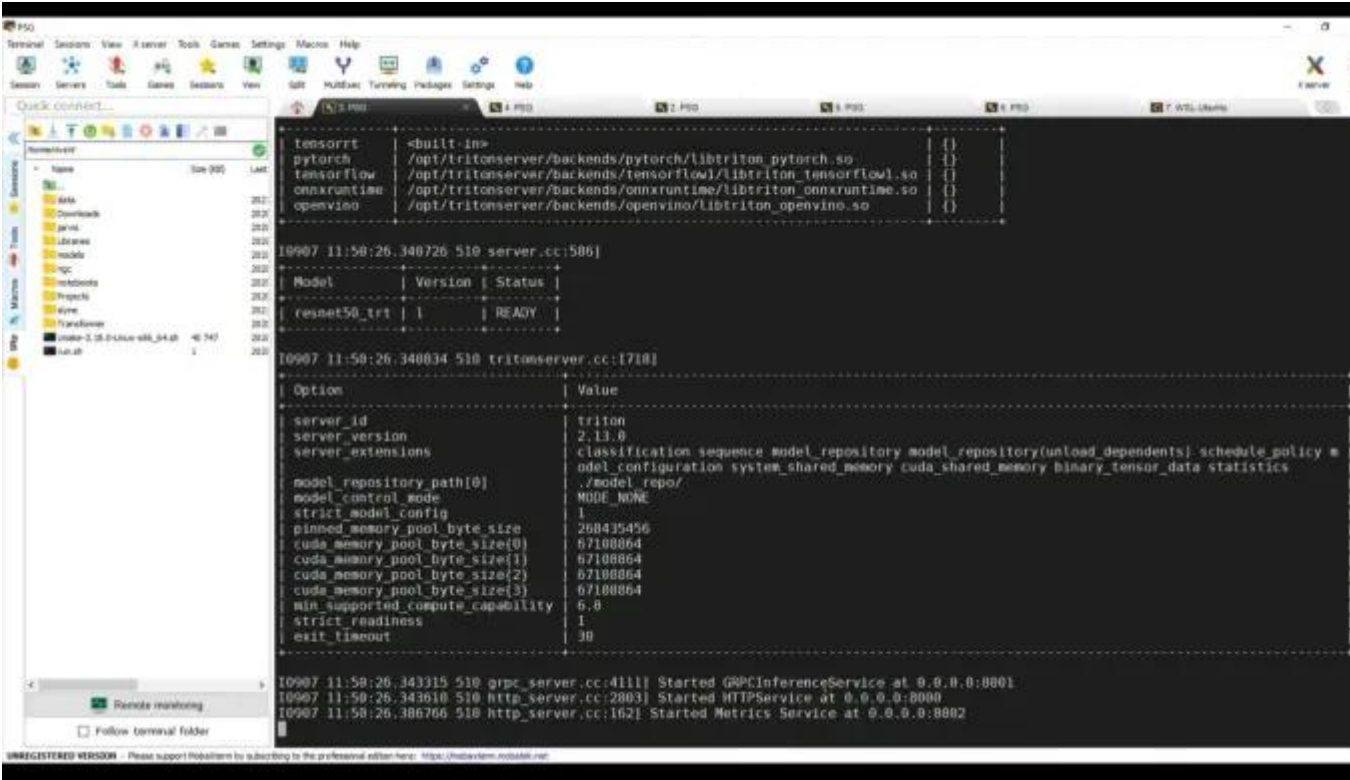
标签文件：

针对分类模型，标签自动转为标签文件中的标签名。

启动 Triton 服务：

启动 Triton 服务时，需要指定仓库目录。

启动成功的输出：



The screenshot shows a terminal window with the following content:

```
tensorrt <built-in>
pytorch /opt/tritonserver/backends/pytorch/libtriton_pytorch.so
tensorflow /opt/tritonserver/backends/tensorflow/libtriton_tensorflow.so
onnxruntime /opt/tritonserver/backends/onnxruntime/libtriton_onnxruntime.so
openvino /opt/tritonserver/backends/openvino/libtriton_openvino.so

10907 11:50:26.340726 510 server.cc:506]

Model Version Status
-----
resnet50_trt 1 READY

10907 11:50:26.340834 510 tritonserver.cc:1718]

Option Value
-----
server_id triton
server_version 2.13.0
server_extensions classification sequence model_repository model_repository(unload_dependents) schedule_policy model_configuration system_shared_memory cuda_shared_memory binary_tensor_data statistics
model_repository_path[0] ./model_repo/
model_control_mode MODE_NONE
strict_model_config 1
pinned_memory_pool_byte_size 268435456
cuda_memory_pool_byte_size[0] 67108864
cuda_memory_pool_byte_size[1] 67108864
cuda_memory_pool_byte_size[2] 67108864
cuda_memory_pool_byte_size[3] 67108864
min_supported_compute_capability 6.0
strict_readiness 1
exit_timeout 30

10907 11:50:26.343315 510 grpc_server.cc:4111] Started GRPCInferenceService at 0.0.0.0:8001
10907 11:50:26.343610 510 http_server.cc:2803] Started HTTPService at 0.0.0.0:8000
10907 11:50:26.386766 510 http_server.cc:162] Started Metrics Service at 0.0.0.0:8082
```