Model serving Architecture

Congratulations! You passed!

Grade received 100% **To pass** 80% or higher

Go to next item

1. What is the core idea of the TensorFlow Serving Architecture? **1 / 1 point**

- (•) The servable
- ( ) The loader
- ( ) The manager
- ( ) The source

Correct
Well done! The servable is the central abstraction in TF-Serving. Clients use these underlying objects to perform computation.

2. True or False: Triton Inference Server simplifies deployment since it is compatible with trained AI models from any framework. **1 / 1 point**

- ( ) False
- (•) True

Correct
Yes! Triton Inference Server allows deployment of models from any framework, from local storage, Google Cloud Platform, or AWS S3.

3. In the TorchServe architecture, where does the actual inference take place? **1 / 1 point**

- (•) Model Workers at the backend
- ( ) Inference endpoints at the frontend
- ( ) Model Store

Correct
That's right! Model Workers are running instances of the model responsible for performing the actual inference. Thus, multiple workers can run simultaneously on Torch Serve.