TensorFlow Serving with LIBXSMM

The TensorFlow Serving framework uses TensorFlow underneath and adds a web-based client/server infrastructure, which can serve requests for inference on an already trained model. The TensorFlow Serving repository (as cloned below) is tracking the master revision of the original Serving framework and it is modified to use a fork of TensorFlow which by itself uses a recent revision of LIBXSMM and the Eigen library (see here).

```
git clone https://github.com/hfp/tensorflow-serving.git
```

It is recommended to use a recent GNU Compiler Collection to build TensorFlow (v5.1 and later). With any recent Bazel version, the desired compiler version can be added to the environment:

```
export PATH=/path/to/gcc/bin:${PATH}
export LD_LIBRARY_PATH=/path/to/gcc/lib64:/path/to/gcc/lib:${LD_LIBRARY_PATH}
export LIBRARY_PATH=/path/to/gcc/lib64:${LIBRARY_PATH}
```

To build the Serving framework execute the following command:

```
bazel build --verbose_failures -c opt --cxxopt=-D_GLIBCXX_USE_CXX11_ABI=0 \
  --copt=-O2 --copt=-fopenmp-simd --copt=-DLIBXSMM_OPENMP_SIMD \
  --define tensorflow_xsmm=1 --define tensorflow_xsmm_convolutions=1 \
  --define tensorflow_xsmm_backward_convolutions=1 \
  --copt=-mfma --copt=-mavx2 \
  --action_env TF_REVISION="master" \
  tensorflow_serving/...
```

If specific target flags are desired (instead of `-mfma -mavx2`), please refer to the TensorFlow with LIBXSMM document. This document can be referred for more details in general.