Building `llama.cpp` on Windows with Vulkan (GPU)

Step 0 — Environment Setup

Open **x64 Native Tools Command Prompt for Visual Studio 2019**. Set CPU/GPU backend flags (can also pass directly to CMake):

```
set GGML_NATIVE=OFF
set GGML_AVX=ON
set GGML_AVX2=OFF
set GGML_F16C=ON
set GGML_FMA=OFF
```

Verify your tools:

```
where cl
cl
where ninja
ninja --version
```

Navigate to project directory:

```
cd /d C:\Users\Admin\source\llama.cpp
```

Step 0.5 — Cleanup / Recovery

Before configuring or switching generators:

```
rmdir /s /q build-vulkan
set CONDA_PREFIX=
set CMAKE_PREFIX_PATH=
set PATH=%PATH:C:\Users\Admin\miniconda3\Library\bin;=%
```

Step 1 — Configure CMake for GPU

One-line configuration without CURL:

```
cmake -S . -B build-vulkan -G "Ninja" -DCMAKE_BUILD_TYPE=Release ^
-DGGML_VULKAN=ON -DGGML_NATIVE=OFF -DGGML_AVX=ON -DGGML_AVX2=OFF ^
-DGGML_F16C=ON -DGGML_FMA=OFF -DJSON_BUNDLED=ON
```

Notes:

- `-DGGML_VULKAN=ON` enables GPU support.
- `-DJSON_BUNDLED=ON` avoids conflicts with system or conda-installed nlohmann::json.
- CURL is not required for basic operation; omit it to avoid build errors.

Step 2 — Build with Ninja

```
cmake --build build-vulkan
```

Expected build time on Ivy Bridge 4c/8t with Vulkan: 30–60 seconds.

Step 3 — Run the server with GPU directly (or port backend)

```
set MODEL=C:\path\to\your\model.gguf
.\build-vulkan\bin\llama-server.exe -m "%MODEL%" --port 8080 --threads 8
```

Base URL for LM Studio: `http://127.0.0.1:8080/v1`

Troubleshooting

- **nlohmann::json LNK2019 errors**: Mixed json ABI versions (v3_11_2 vs v3_12_0). Solution: clean build folder, clear `CMAKE_PREFIX_PATH`, and rebuild with `-DJSON_BUNDLED=ON`.
- **CMake cannot find CURL**: Not critical; set `-DLLAMA_CURL=OFF` or just omit CURL as shown.
- **Ninja errors regarding compiler**: Ensure you are in x64 Native Tools prompt and that `build-vulkan` is freshly created.

One-line GPU Ninja build (copy/paste)

```
cd /d C:\Users\Admin\source\llama.cpp && rmdir /s /q build-vulkan 2>nul &&
set CONDA_PREFIX= && set CMAKE_PREFIX_PATH= && set PATH=%PATH:C:\Users\Admin\
    miniconda3\Library\bin;=% &&
cmake -S . -B build-vulkan -G "Ninja" -DCMAKE_BUILD_TYPE=Release ^
-DGGML_VULKAN=ON -DGGML_NATIVE=OFF -DGGML_AVX=ON -DGGML_AVX2=OFF ^
-DGGML_F16C=ON -DGGML_FMA=OFF -DJSON_BUNDLED=ON &&
cmake --build build-vulkan
```