Nvidia GPU Setup

Prerequisites / Assumptions

This guide will assume that you are using `x86_64` processor with Linux Ubuntu 24.04 LTS installed on it. Furthermore it is assumed that you have a Nvidia GPU with a Compute Capability of 5.0 or above.

These assumptions are curcial because the LLM-Backend (Ollama) makes those assumptions and more.

Setting up the GPU

1. **Install Nvidia GPU driver**

    1. Installing can happen in two major ways.

        a) Using the GUI

        > If you are on a Desktop-Setup use this method. It is the most straightforward way

        b) Using the Terminal

        > For a server or headless setup

        - List all available hardware drivers

            ```
            sudo ubuntu-drivers list
            ```

        - Install drivers

            - Using automatic detection (*recommended way*)

                ```
                sudo ubuntu-drivers install
                ```

            - Using manual selection

```
        # Let's Assume you want to install the 535 nvidia dirver
        sudo ubuntu-drivers install nvidia:535
        ```

    2. Verifying installation

        ```
        # You may need to reboot your system before changes take affect using sudo reboot
        nvidia-smi
        ```

2. **Installing Cuda Libraries**

    Installing Cuda libraries is crucial since they define *how* your GPU is supposed to perform calculations.

    1. Installing `nvidia-cudo-toolkit`

        a) Easiest / Recommended way

        Install the `nvidia-cuda-toolkit`. This package is the all-in-one bundle. It'll contain unnecessary bloat, and will be most likely not the latest version.

        ```
        sudo apt install nvidia-cuda-toolkit
        ```

        b) From Nvidia

        Installing the Nvidia toolkit this way get's you the latest version directly from nvidia. This installation requires access to GitHub like services which makes them less reliable in Mainland China.

        ```
        # Install the cuda-keyring package
        wget https://developer.download.nvidia.com/compute/cuda/repos/<distro>/<arch>/cuda-keyring_1.1-1_all.deb

        sudo dpkg -i cuda-keyring_1.1-1_all.deb

        dpkg -i cuda-keyring_1.1-1_all.deb

        apt update
        ```

        The following instruction **must** be executed separately

```
apt install cuda-toolkit
```

nvidia-gds is optional but it is strongly recommended to install it, since it greatly enhances GPU performance.

```
apt install nvidia-gds

reboot
```

2. Verifying Installation

```
nvcc --version
```

3. **Installing nvidia container toolkit**

   The nvidia_container_toolkit allows docker to make use of your GPU.

   1. Installing nvidia_container_toolkit

      - Configure production repository:

      ```
      curl -fsSL https://nvidia.github.io/libnvidia-container/gpgkey | sudo gpg --dearmor -o /usr/share/keyrings/nvidia-container-toolkit-keyring.gpg \
      && curl -s -L https://nvidia.github.io/libnvidia-container/stable/deb/nvidia-container-toolkit.list | \
      sed 's#deb https://#deb [signed-by=/usr/share/keyrings/nvidia-container-toolkit-keyring.gpg] https://#g' | \
      sudo tee /etc/apt/sources.list.d/nvidia-container-toolkit.list
      ```

      - Update the packages list from the repository:

      ```
      sudo apt-get update
      ```

      - Install the NVIDIA Container Toolkit packages:

      ```
      export NVIDIA_CONTAINER_TOOLKIT_VERSION=1.17.8-1
      sudo apt-get install -y \
      nvidia-container-toolkit=${NVIDIA_CONTAINER_TOOLKIT_VERSION} \
      nvidia-container-toolkit-base=${NVIDIA_CONTAINER_TOOLKIT_VERSION} \
      libnvidia-container-
      ```

```
tools=${NVIDIA_CONTAINER_TOOLKIT_VERSION} \
libnvidia-container1=${NVIDIA_CONTAINER_TOOLKIT_VERSION}
```

- Configuring Docker

```
sudo nvidia-ctk runtime configure --runtime=docker
```

2. Verifying installation

```
# For the free world
sudo docker run --rm --runtime=nvidia --gpus all ubuntu nvidia-smi

# For mainland china
sudo docker run --rm --runtime=nvidia --gpus all docker.1ms.run/ubuntu
nvidia-smi
```

Sources

- Installing Nvidia Drivers
- Installing Nvidia Cuda Toolkit via package manager
- Installing Nvidia Cuda Toolkit via Nvidia repository
- What is Nvidia GDS
- Installing Nvidia Container Toolkit