

Docker Ollama avec support Nvidia

A utiliser sous Ubuntu 22.4 Le port du service est 3000

Drivers NVIDIA

```
lspci | grep -i nvidia
apt install ubuntu-drivers-common -y
ubuntu-drivers install
```

Drivers NVIDIA Cuda

```
wget
https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64/c
uda-keyring_1.1-1_all.deb
dpkg -i cuda-keyring_1.1-1_all.deb
apt update
apt-get -y install cuda-toolkit-12-6
```

NVIDIA Container Toolkit

```
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
```

```
sed -i -e '/experimental/ s/^#//g' /etc/apt/sources.list.d/nvidia-container-
toolkit.list
apt update
apt-get install -y nvidia-container-toolkit
nvidia-ctk runtime configure --runtime=docker
systemctl restart docker
```

Test

```
nvidia-smi
docker run --rm -it --gpus all nvcr.io/nvidia/pytorch:22.03-py3
```

Docker Compose

```
services:
  ollama:
    image: ollama/ollama:latest
    hostname: ollama
    ports:
      - "11434:11434"
    volumes:
      - ./models:/root/.ollama/models
    networks:
      - genai-network
    deploy:
      resources:
        reservations:
          devices:
            - driver: nvidia
              count: all
              capabilities: [gpu]
  open-webui:
    image: ghcr.io/open-webui/open-webui:main
    ports:
      - "3000:8080"
    environment:
      - OLLAMA_BASE_URL=http://ollama:11434
    volumes:
      - ./backend/data:/app/backend/data
    networks:
      - genai-network
networks:
  genai-network:
    driver: bridge
    name: genai-network
```

Sources

- (pas relevant) <https://dev.to/thenjdevopsguy/using-nvidia-gpus-with-docker-in-5-minutes-386g>
- <https://dev.to/thenjdevopsguy/using-nvidia-gpus-with-docker-in-5-minutes-386g>
- <https://medium.com/@srpillai/how-to-run-ollama-locally-on-gpu-with-docker-a1ebabe451e0>
- <https://github.com/sujithrpillai/ollama/blob/main/ollama/docker-compose.yml>

From:
<https://wiki.makeitsimple.be/> - **makeITsimple wiki**

Permanent link:
<https://wiki.makeitsimple.be/doku.php?id=linux:docker-nvidia-ollama>

Last update: **2025/02/25 15:48**



