




# NCHC LLM Bootcamp

Cliff Chiu, Solution Architect | June 18, 2025



# 在今天上課之前，請有到Huggingface取得模型授權

## Llama-3.1-8B-Instruct

 **Hugging Face**

Models

Datasets

Spaces

Community

Docs

Pricing

Log In

Sign Up

meta-llama/

**Llama-3.1-8B-Instruct**

like

4.11k

Follow

Meta Llama

48.3k

Text Generation

Transformers

Safetensors

PyTorch

8 languages

llama

facebook

meta

llama-3

conversational

text-generation-inference

arxiv:2204.05149

License: llama3.1

Model card

Files and versions

xet

Community

235

Train

Deploy

Use this model

You need to agree to share your contact information to access this model

The information you provide will be collected, stored, processed and shared in accordance with the [Meta Privacy Policy](#).

LLAMA 3.1 COMMUNITY LICENSE AGREEMENT

Llama 3.1 Version Release Date: July 23, 2024

"Agreement" means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.

"Documentation" means the specifications, manuals and documentation accompanying Llama 3.1 distributed by Meta at <https://llama.meta.com/doc/overview...>

Log in

or

Sign Up

to review the conditions and access this model content.

Downloads last month

5,386,403

Safetensors

Model size

8.03B params

Tensor type

BF16

Chat template

Files info

Inference Providers

NEW


Fireworks

+4

Text Generation

Examples

Input a message to start chatting with meta-llama/Llama-3.1-8B-Instruct.



## 事前準備工作

1. 使用TWCC平台建立 NeMo-24.12:latest 開發環境
2. 下載本次教材 **LLM-Tutorial – Github**

```
git clone https://github.com/wcks13589/LLM-Tutorial.git  
cd LLM-Tutorial
```

3. 登入 Huggingface 取得 **HF\_TOKEN**

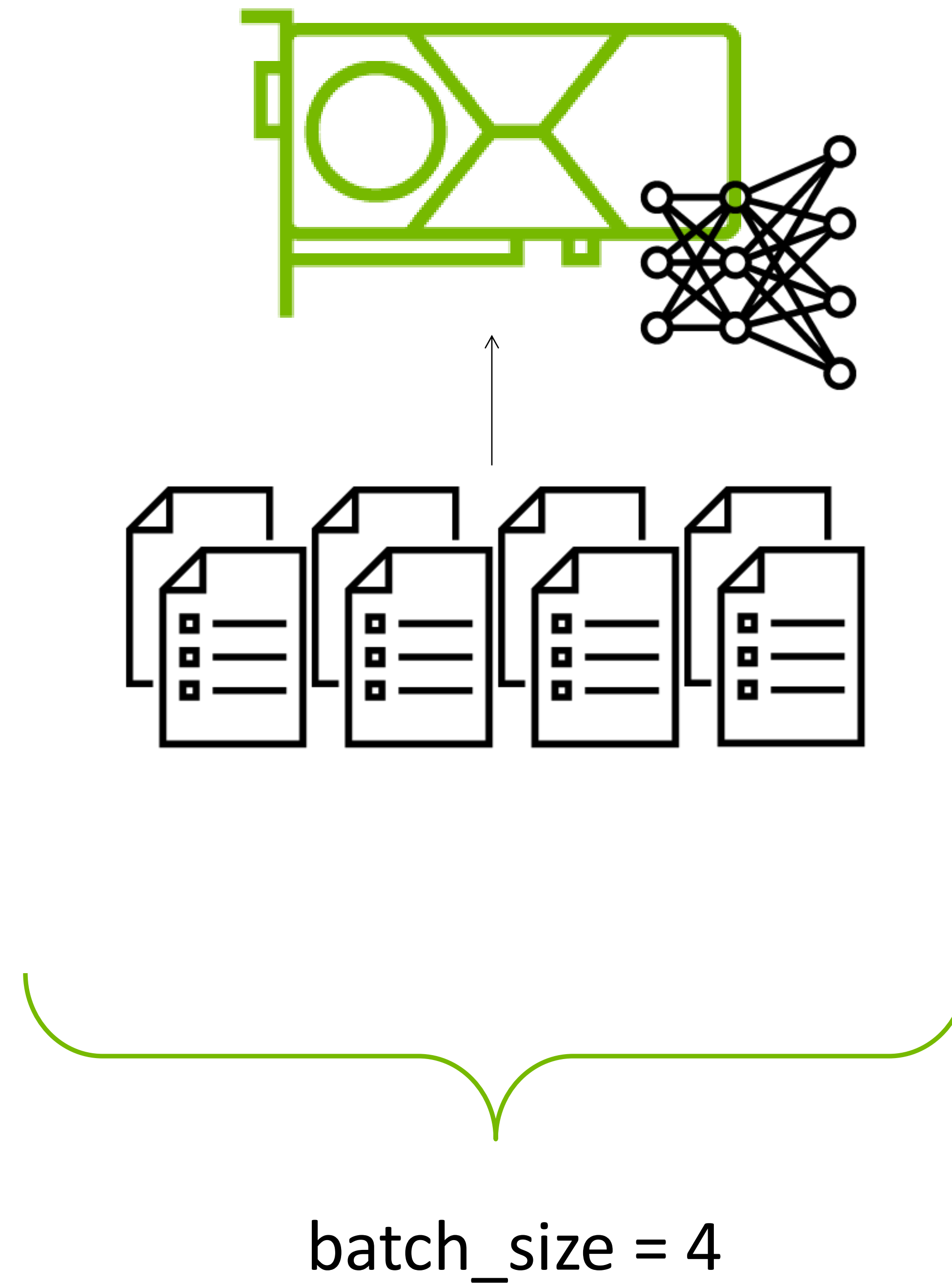
```
export HF_TOKEN="your_hf_token"  
huggingface-cli login --token $HF_TOKEN
```

4. 下載 **Llama3.1-8B-Instruct** 模型到 ~/LLM-Tutorial/

```
huggingface-cli download meta-llama/Llama-3.1-8B-Instruct \  
  --local-dir Llama-3.1-8B-Instruct \  
  --exclude original/
```

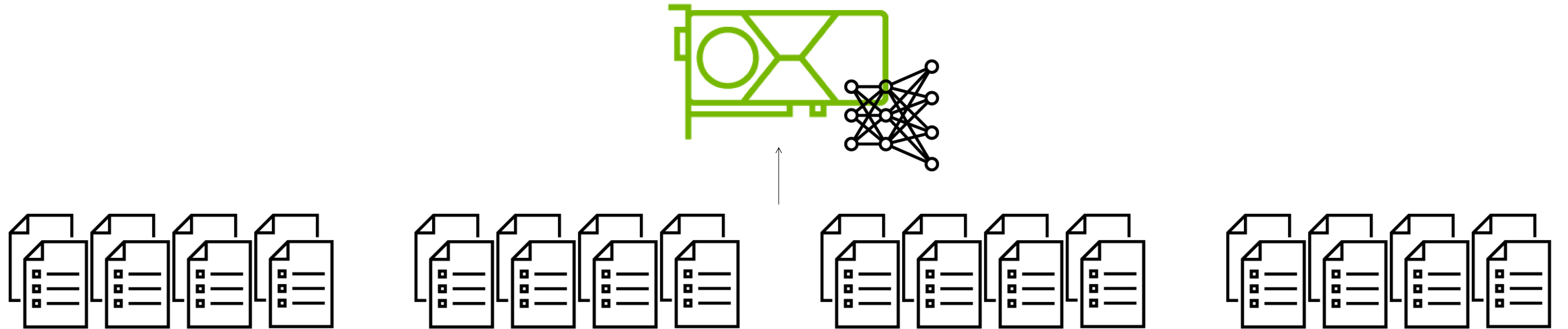
# Batch size

GPU



# Batch size

GPU



batch\_size = 16



# Micro\_batch\_size vs Global\_batch\_size

GPU 1



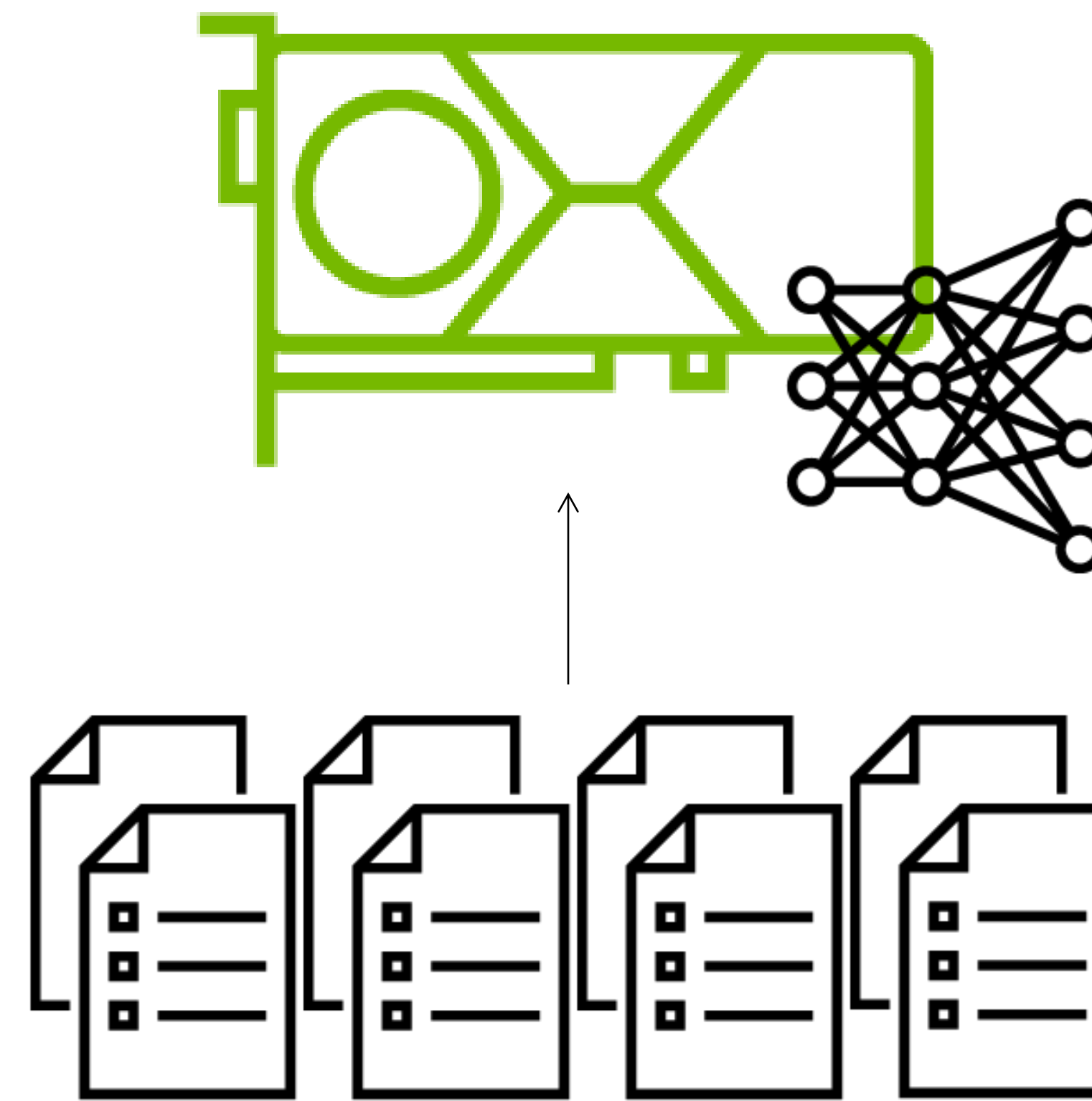
Micro\_batch\_size = 4

GPU 2



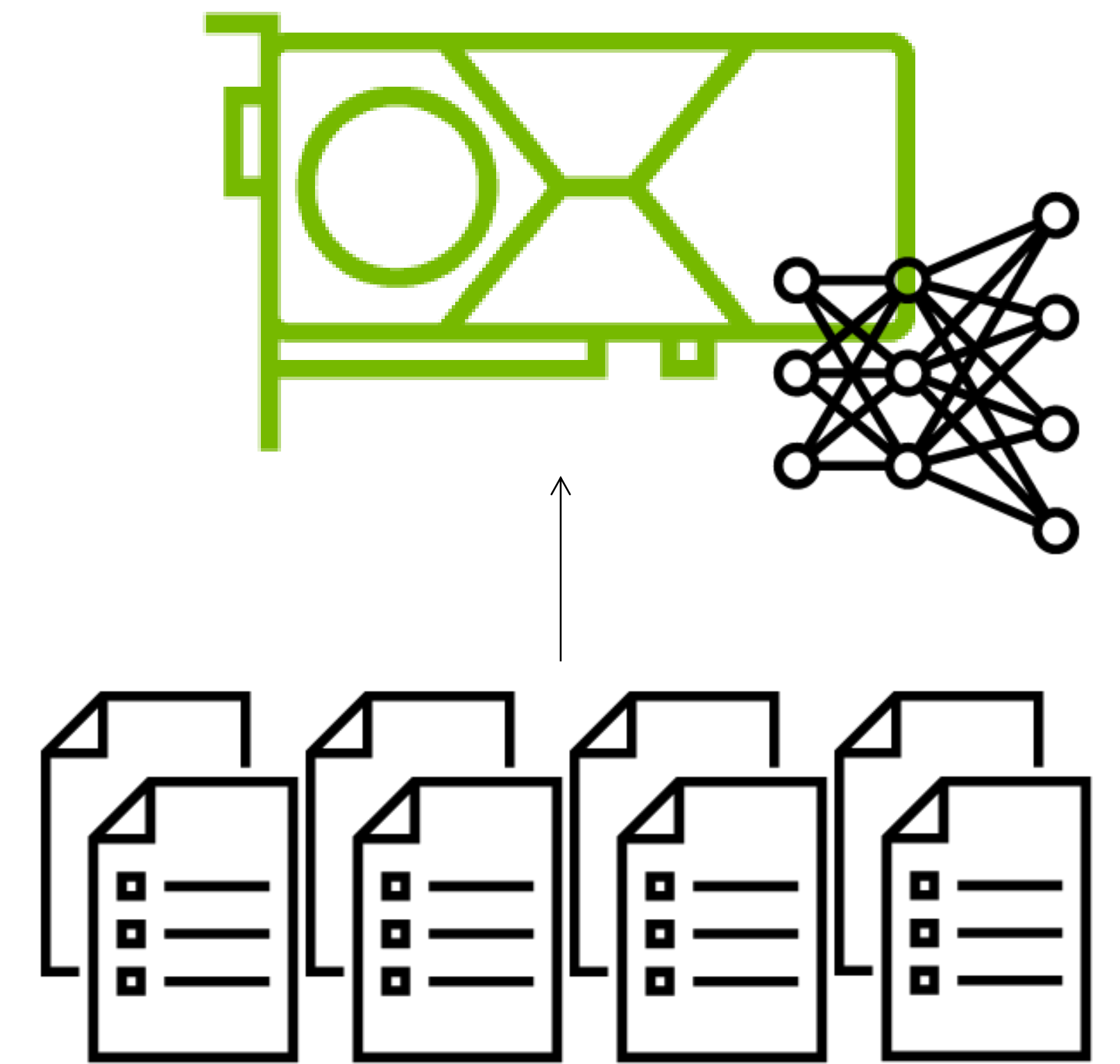
Micro\_batch\_size = 4

GPU 3



Micro\_batch\_size = 4

GPU 4

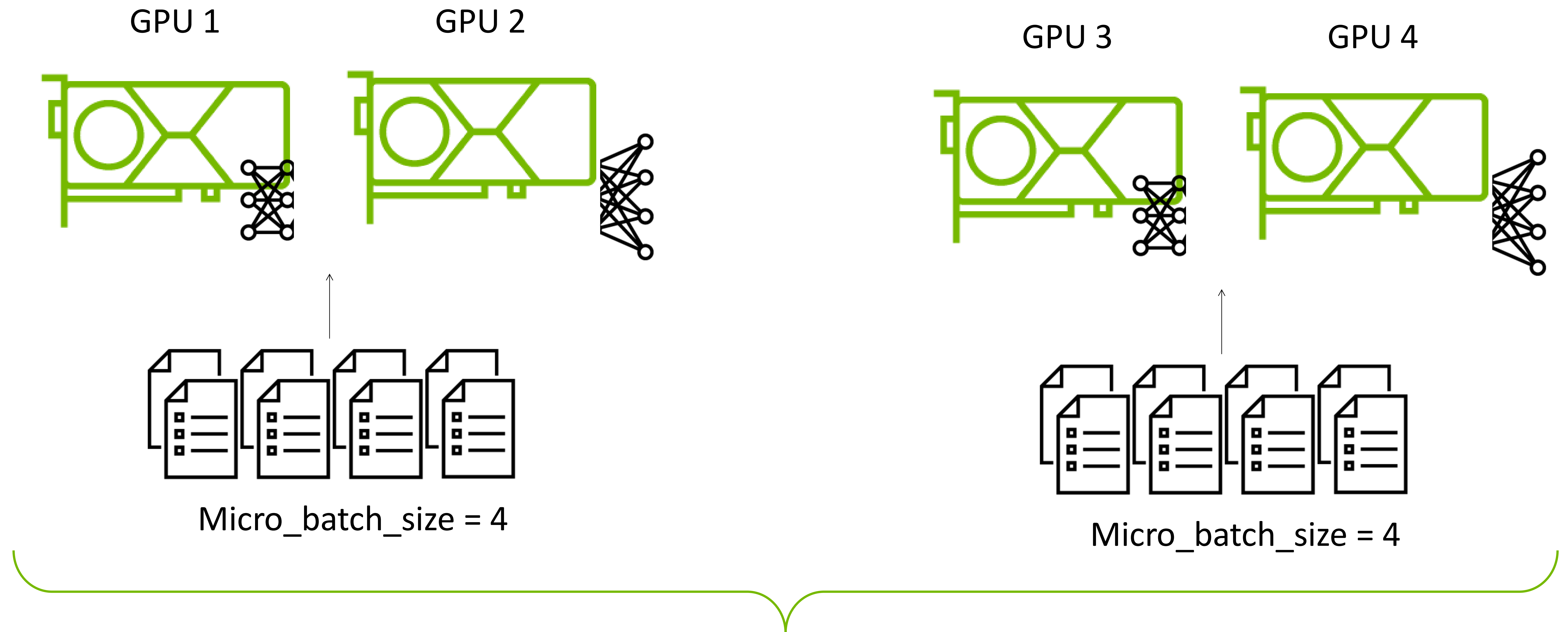


Micro\_batch\_size = 4

Tensor\_parallel\_size = 1  
Pipeline\_parallel\_size = 1

Global\_batch\_size = 16 = 4 x 4 (Micro\_batch\_size \* Data\_parallel\_size)

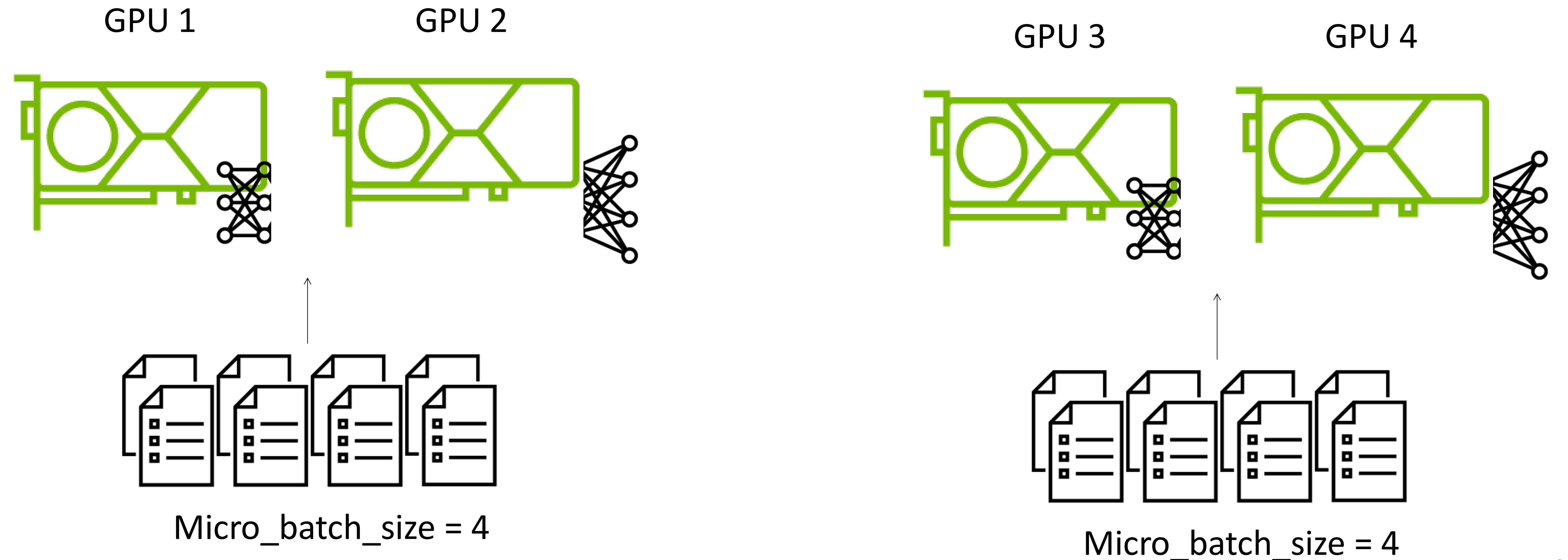
# Micro\_batch\_size vs Global\_batch\_size



Tensor\_parallel\_size = 2  
Pipeline\_parallel\_size = 1

Global\_batch\_size = 8 = 4 x 2 (Micro\_batch\_size \* Data\_parallel\_size)

# Micro\_batch\_size vs Global\_batch\_size



Tensor\_parallel\_size = 2  
Pipeline\_parallel\_size = 1

Global\_batch\_size = 16 = 4 x 2 x 2 (Micro\_batch\_size \* Data\_parallel\_size \* accumulate\_grad\_batches)



## How to set **Max\_steps**

- 資料集 : erhwenkuo/wikinews-zhtw
  - 經過Llama2的Tokenize之後, 共有 **7198329** 個 token
  - Total token = **7198329**
- seq\_length: **2048** (TinyLlama)
- total\_samples =  $7198329 / 2048 \approx$  **3515**
- global\_batch\_size = 4
- 1 epoch =  $3515 / 4 \approx$  **879 steps**