▲□▶ ▲□▶ ▲ □▶ ▲ □▶ □ のへぐ

# Efficient training of Neural Networks

Adrien Aguila--Multner

Enseirb-Matmeca & Inria

March 20, 2024



Artificial Intelligence

Optimisation

High Performance Computing

イロト 不得 トイヨト イヨト

3



▲□▶ ▲□▶ ▲□▶ ▲□▶ ■ ●の00

## Internship topic(s)

- Efficient training of neural networks
- Pipeline parallelism





• Activation checkpointing (Rockmate)

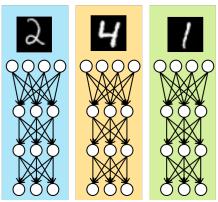
Introduction O Internship topic(

3d parallelism

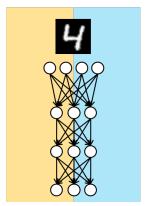
Tools 0 Conclusion o

### Data Parallel vs Model Parallel

## Data Parallel



# Model Parallel



▲□▶ ▲□▶ ▲ □▶ ▲ □▶ □ のへぐ

Internship topic(s)

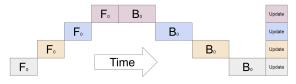
3d parallelism

Tools 0

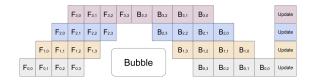
▲□▶ ▲□▶ ▲ 三▶ ▲ 三▶ 三 のへぐ

Conclusion 0

#### **Pipeline Parallel**



Without pipelining



With pipelining

Introduction O Internship topic(s)

3d parallelism

Tools

▲□▶ ▲□▶ ▲ 三▶ ▲ 三▶ 三三 - のへぐ

Conclusion

#### Torch distributed & Torchrun

#### Demo

Introduction O

▲□▶ ▲□▶ ▲ 三▶ ▲ 三▶ 三三 - のへぐ



# Thank you !