

Pipeline Parallelism

Instructor

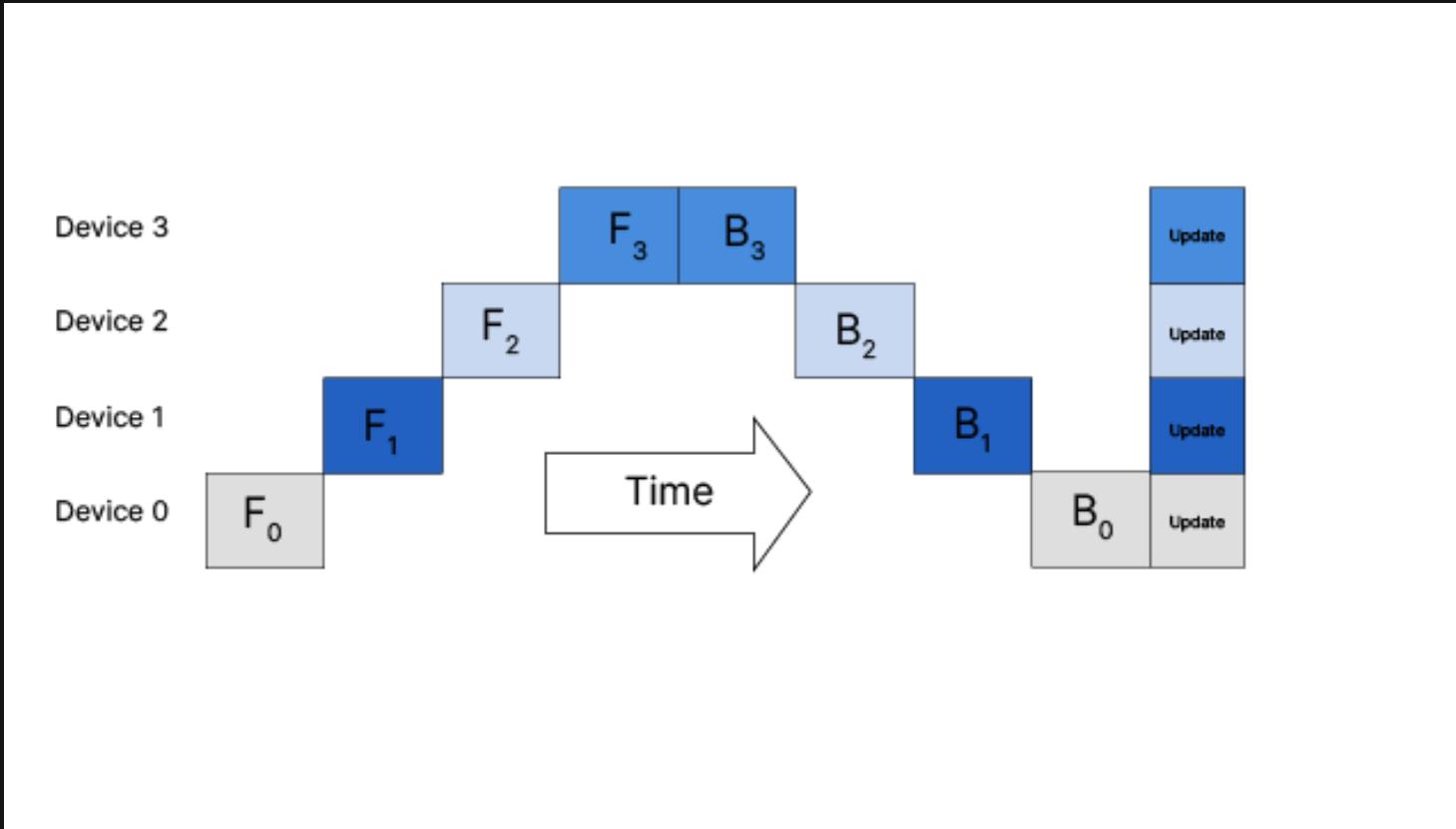
Sourab Mangrulkar

Machine Learning Engineer at

Creator of PEFT

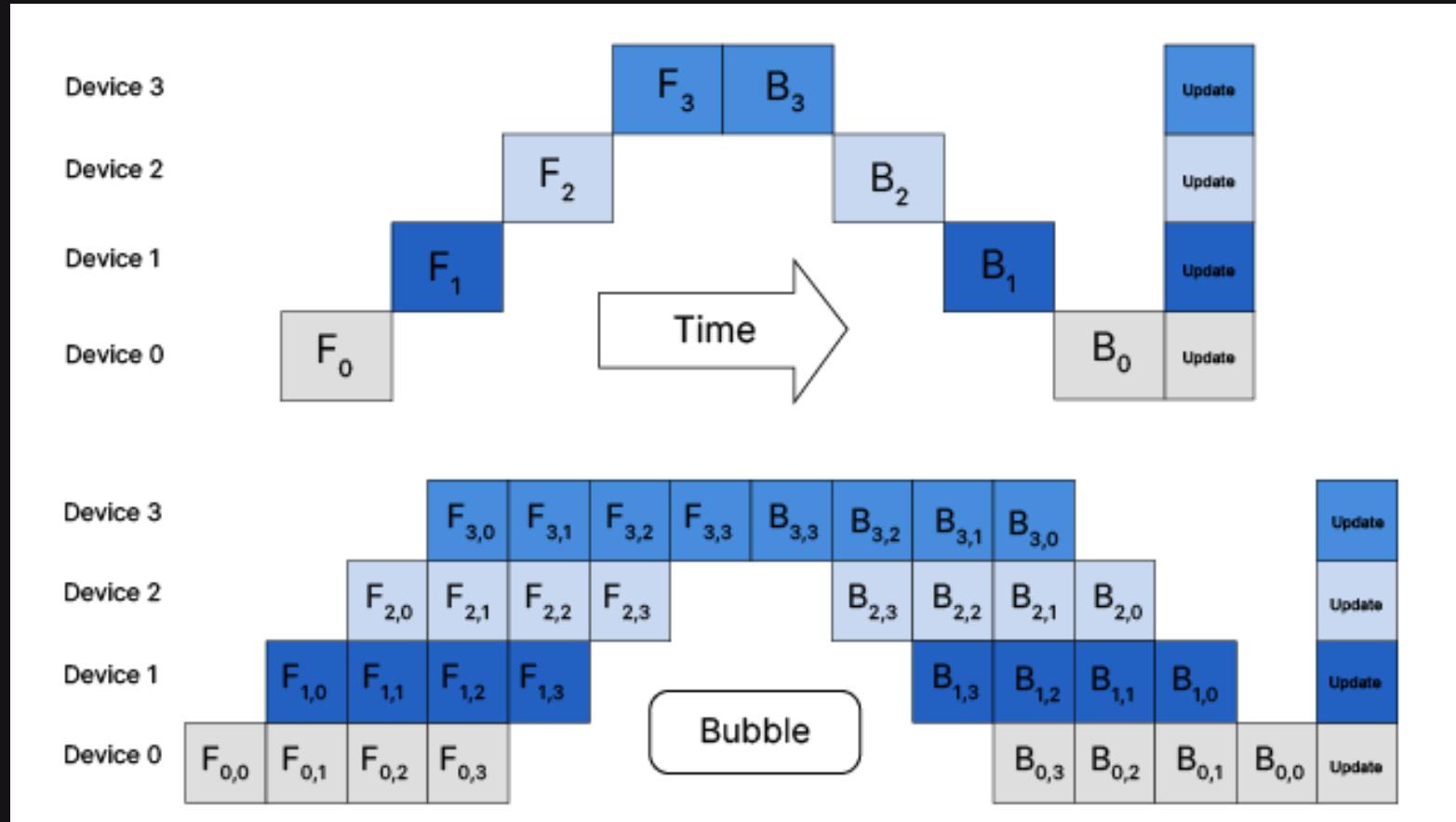


Model Parallelism



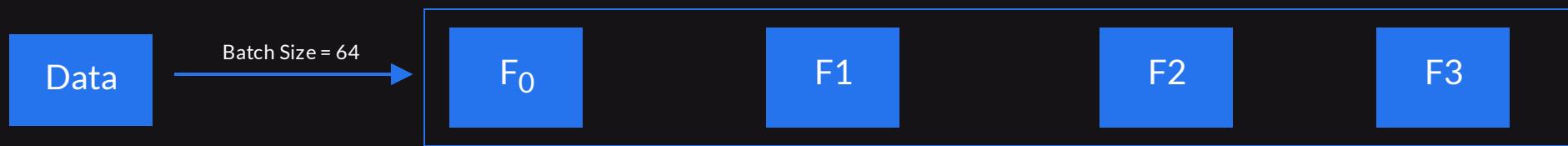
Pipeline Parallelism

Pipeline Parallelism



The Process

Number of GPUs: 4



The Process

GPU 3

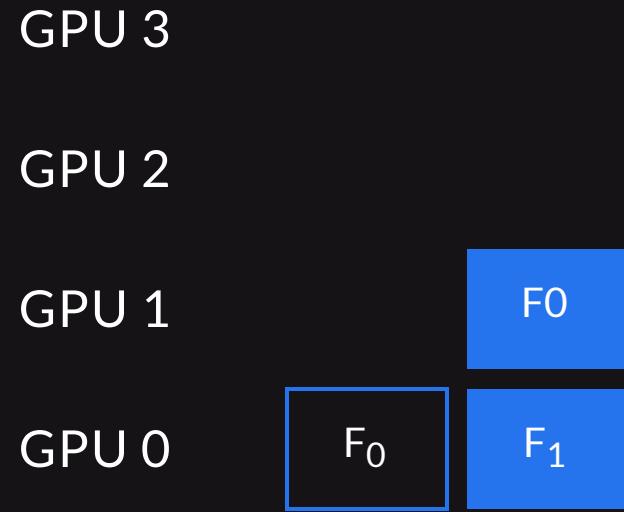
GPU 2

GPU 1

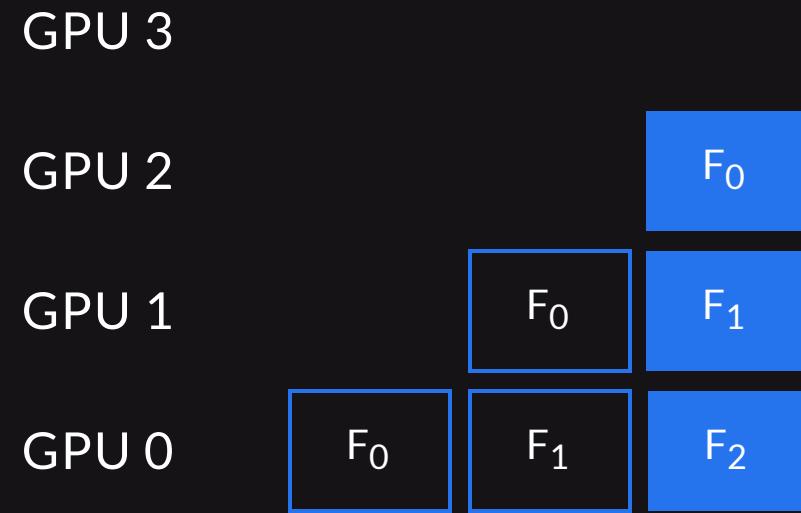
GPU 0

F_0

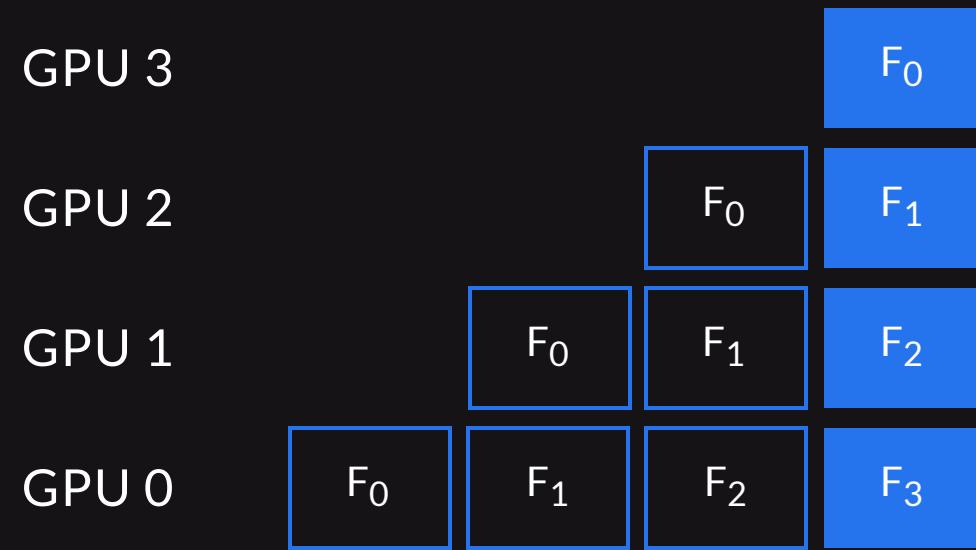
The Process



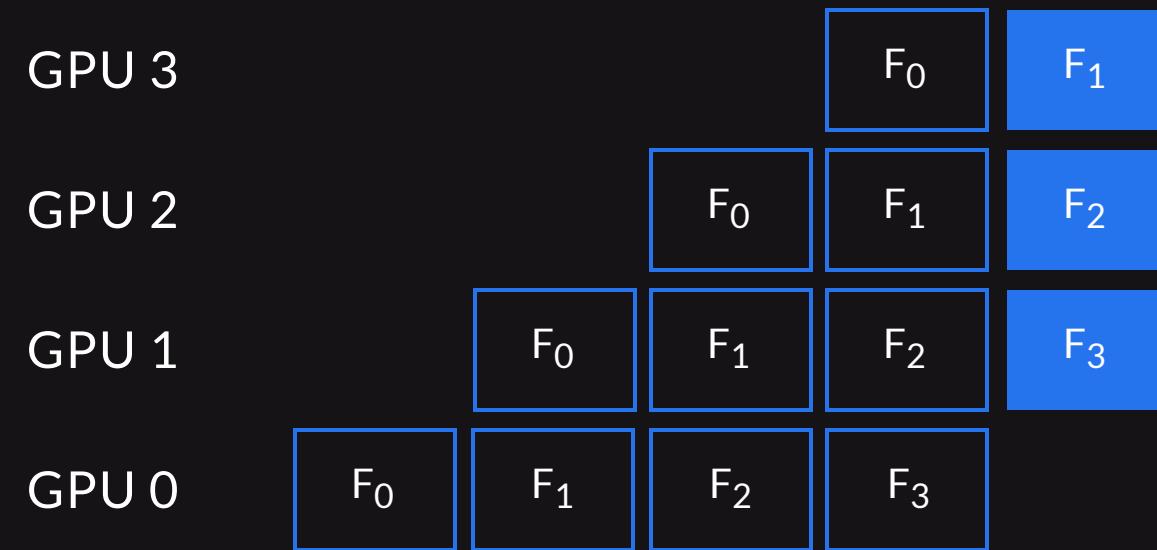
The Process



The Process



The Process



The Process

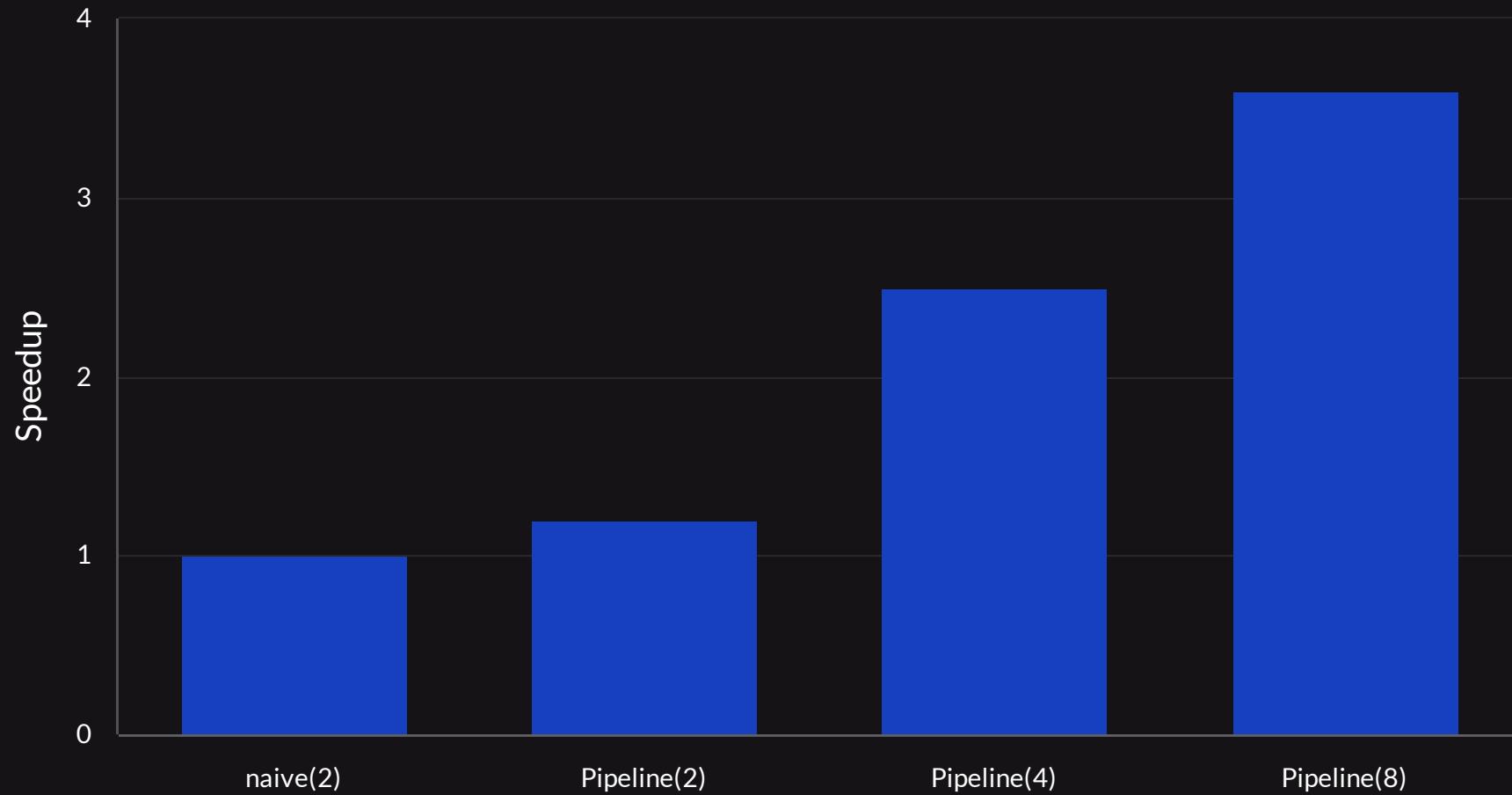


The Process



Model Parallelism Vs Pipeline Parallelism

AmoebaNet-D (4,512)



Cons

- Communication Overhead

Cons

- Communication Overhead
- Underutilisation of GPUs

Thank You
