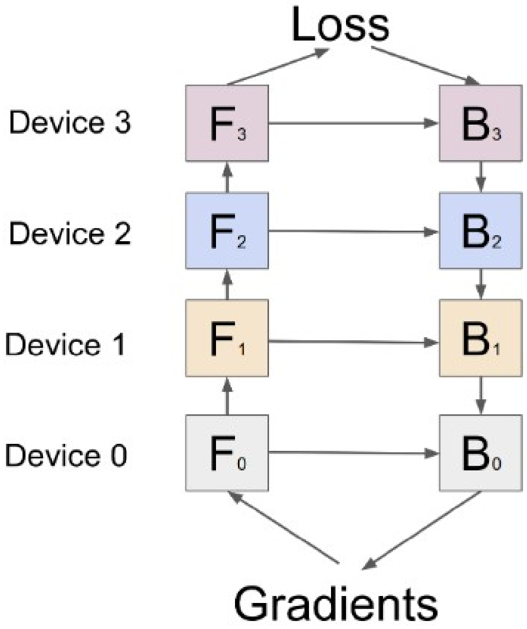
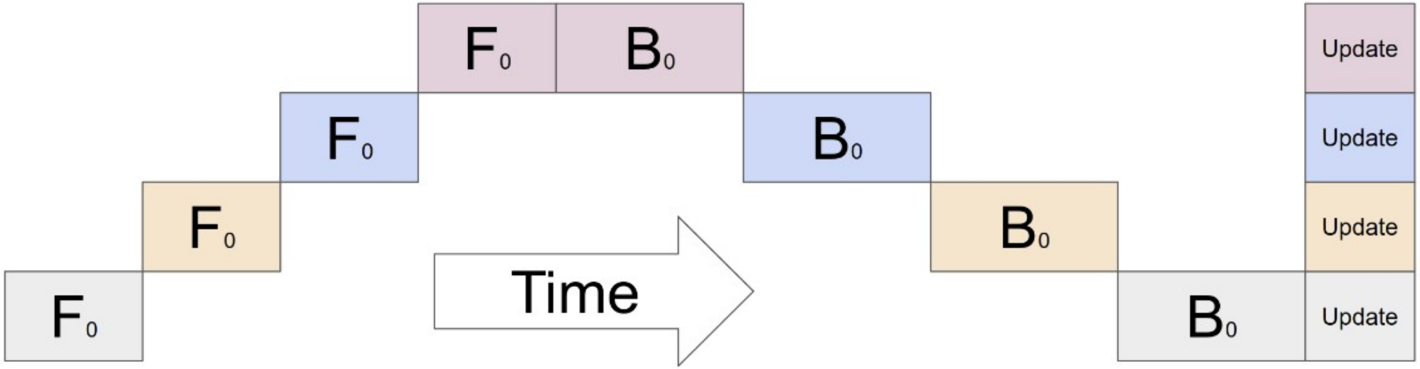
Model-parallel training

Q: What if a model is larger than GPU?





model size: O(N)
throughput: O(1)

Q: Can we go faster?