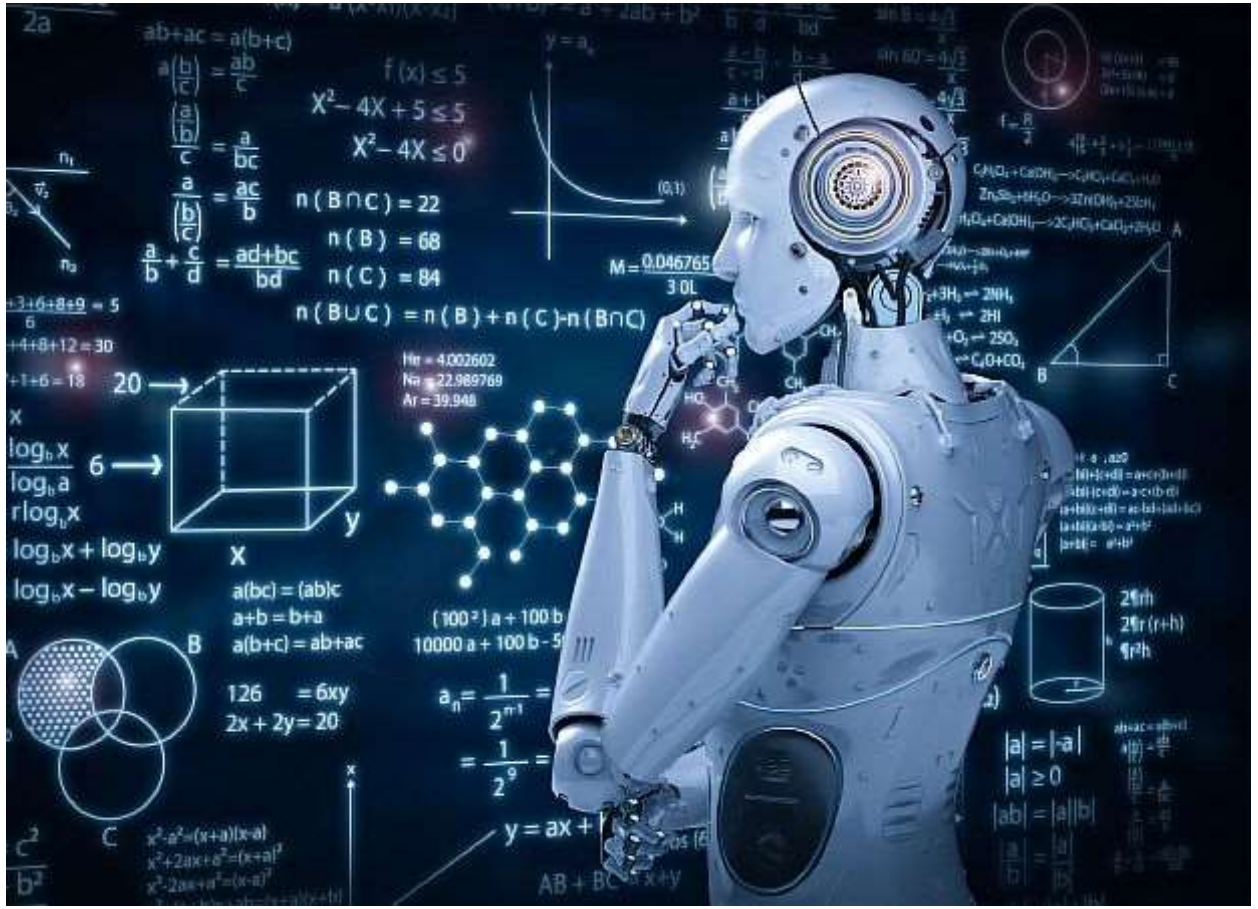
# РАСПРЕДЕЛЁННОЕ ОБУЧЕНИЕ НЕЙРОСЕТЕЙ

Торбахов Тимофей БПМИ211

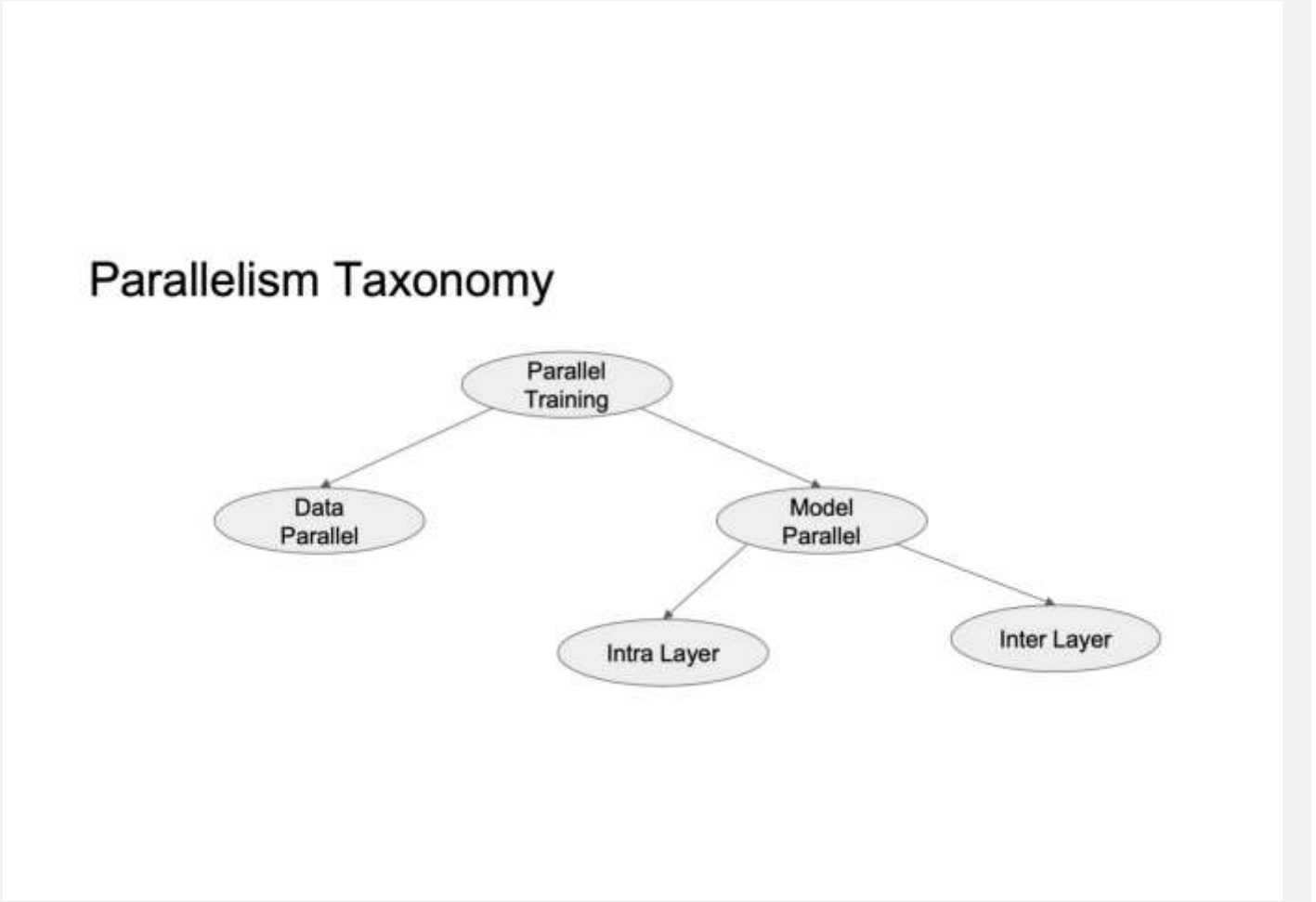



# ЗАЧЕМ?

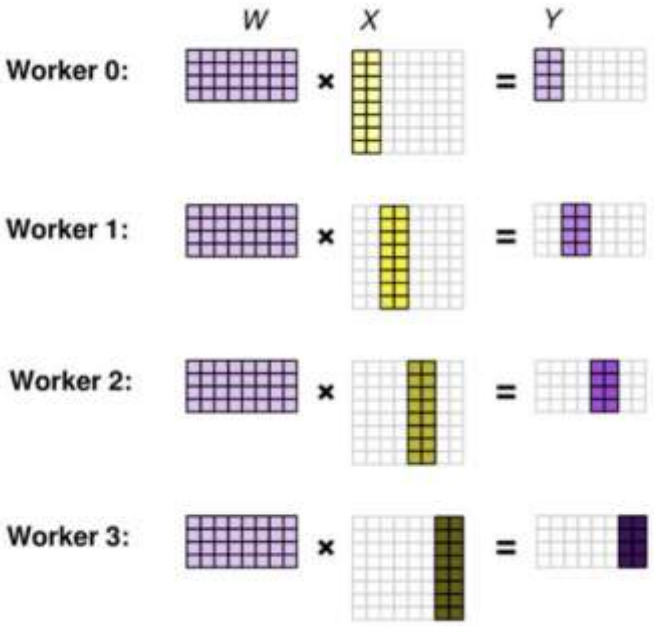
Как и всегда: время и память!

# ПОДХОДЫ К РАСПРЕДЕЛЁННОМУ ОБУЧЕНИЮ

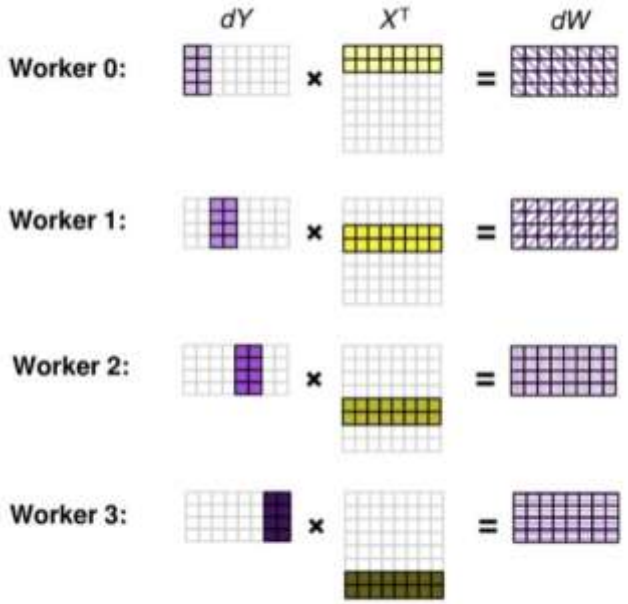
Обычно выделяют два подхода: параллелизм данных и параллелизм моделей, последний, в свою очередь, также делится на tensor и pipeline parallelism



Прямой проход

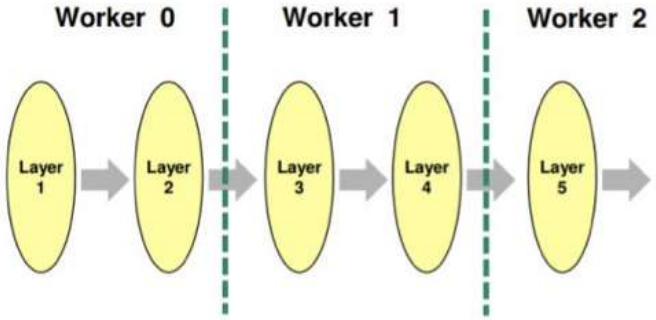



Обратный проход

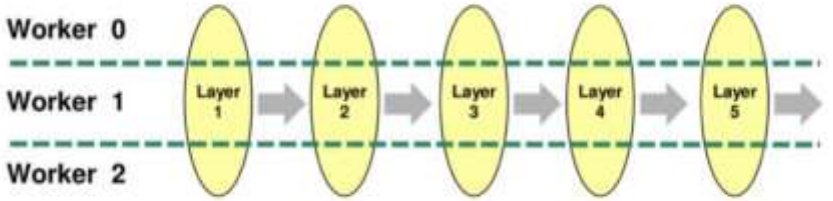



# DATA PARALLELISM

Pipeline parallelism



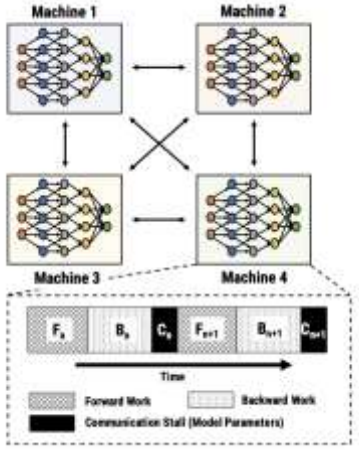

Tensor parallelism



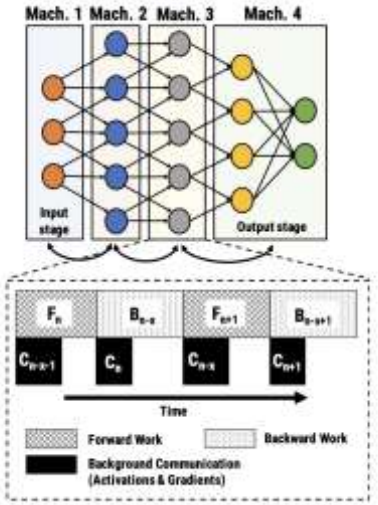

# MODEL PARALLELISM

# PIPEDREAM

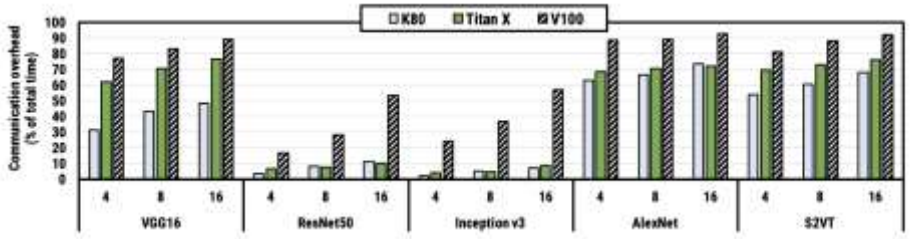
# МОТИВАЦИЯ



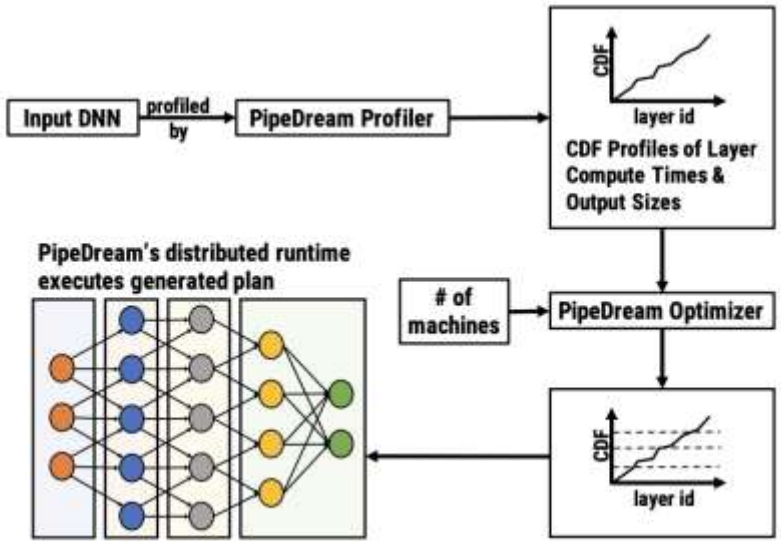

Data parallel

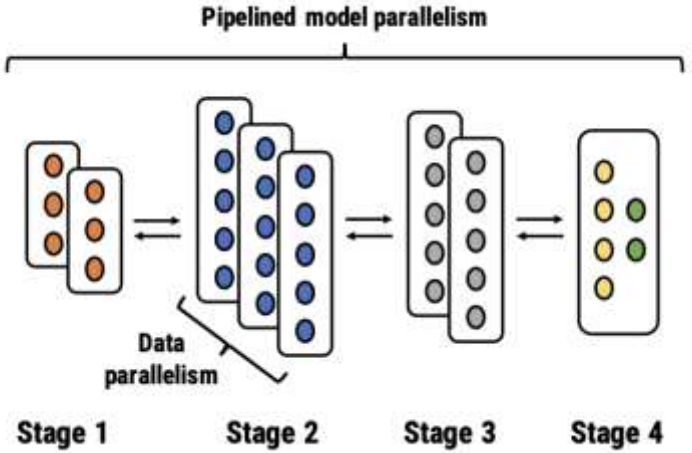



Pipeline parallel

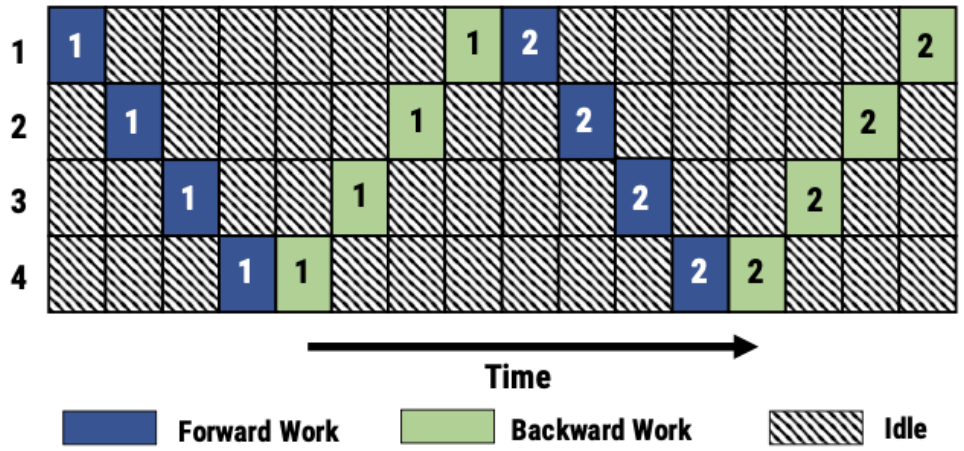


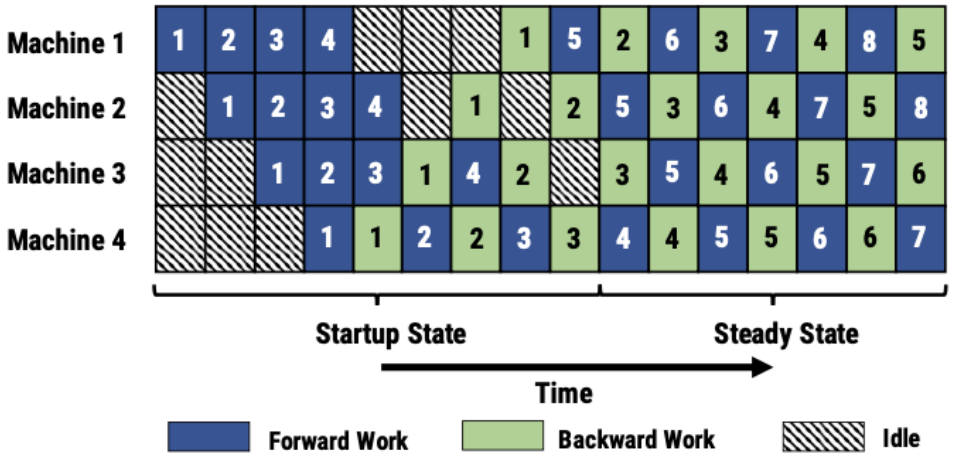

Communication overhead



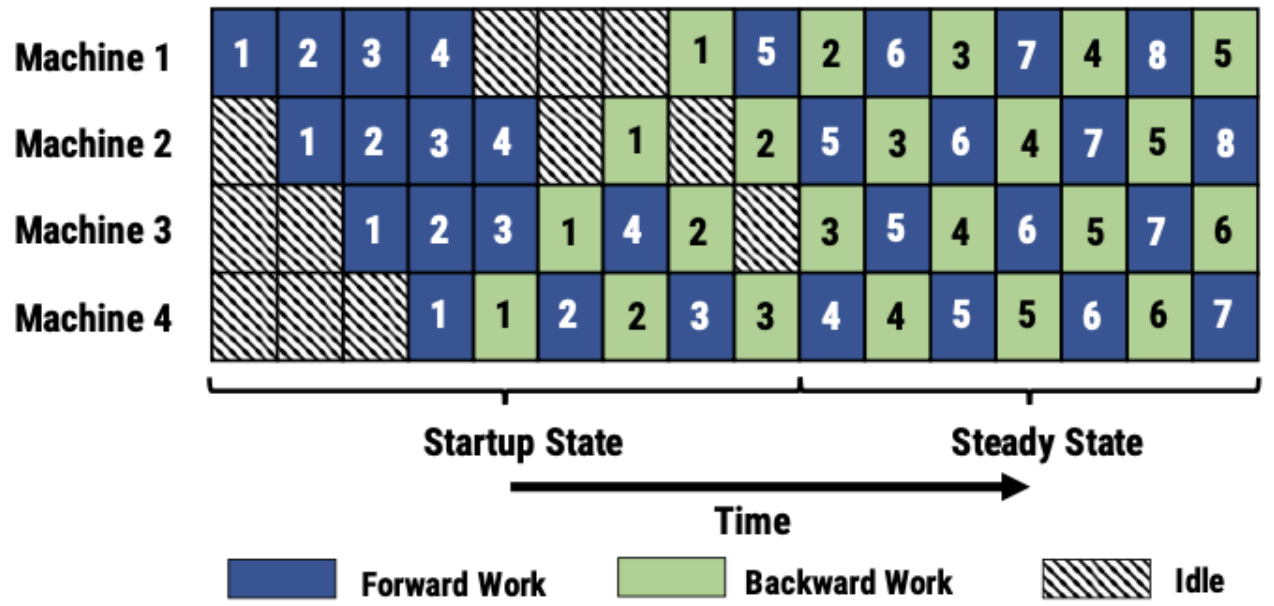

# АВТОМАТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ СТУПЕНЕЙ МЕЖДУ GPU




# ПЛАНИРОВАНИЕ РАБОТЫ

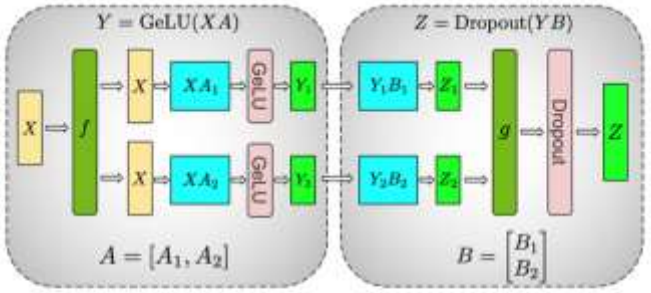
ЭФФЕКТИВНОСТЬ ОБУЧЕНИЯ



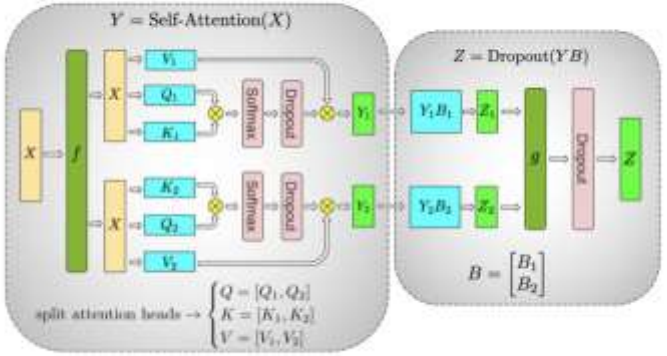
# MEGATRONLM

# МОТИВАЦИЯ

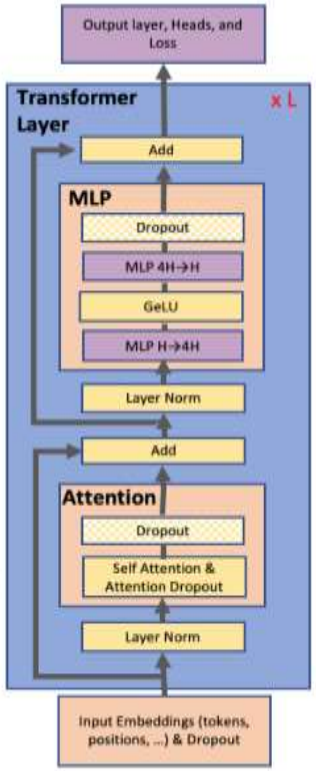
С увеличением размера и сложности языковых моделей, таких как BERT и GPT-2, нейронные сети **приблизились к объему памяти** современных аппаратных ускорителей




(a) MLP




(b) Self-Attention




# МОДЕЛЬНЫЕ ПАРАЛЛЕЛЬНЫЕ ТРАНСФОРМАТОРЫ

# СЛОЖНОСТИ ОБУЧЕНИЯ С БОЛЬШИМ БАТЧОМ

# ЗАЧЕМ?

Для эффективной работы распределённого SGD требуется, чтобы нагрузка на каждого «работника» была велика, что влечёт за собой нетривиальный рост размера минибатча.

# ОПТИМИЗАЦИИ

1. *Когда увеличиваем размера минибатча в k раз, необходимо умножать leaning rate на k*
2. *Чтобы избежать проблем с п.1, необходимо делать «разминку» модели*
3. *Масштабирование функции потерь не эквивалентно масштабированию learning rate*
4. *Применять momentum correction после изменения leraning rate*
5. *Использовать одно перемешивание данных за эпоху*

# ИСТОЧНИКИ ИНФОРМАЦИИ

- Distributed learning and network computing
- Pipedream
- MegatronLM
- Large Minibatch SGD