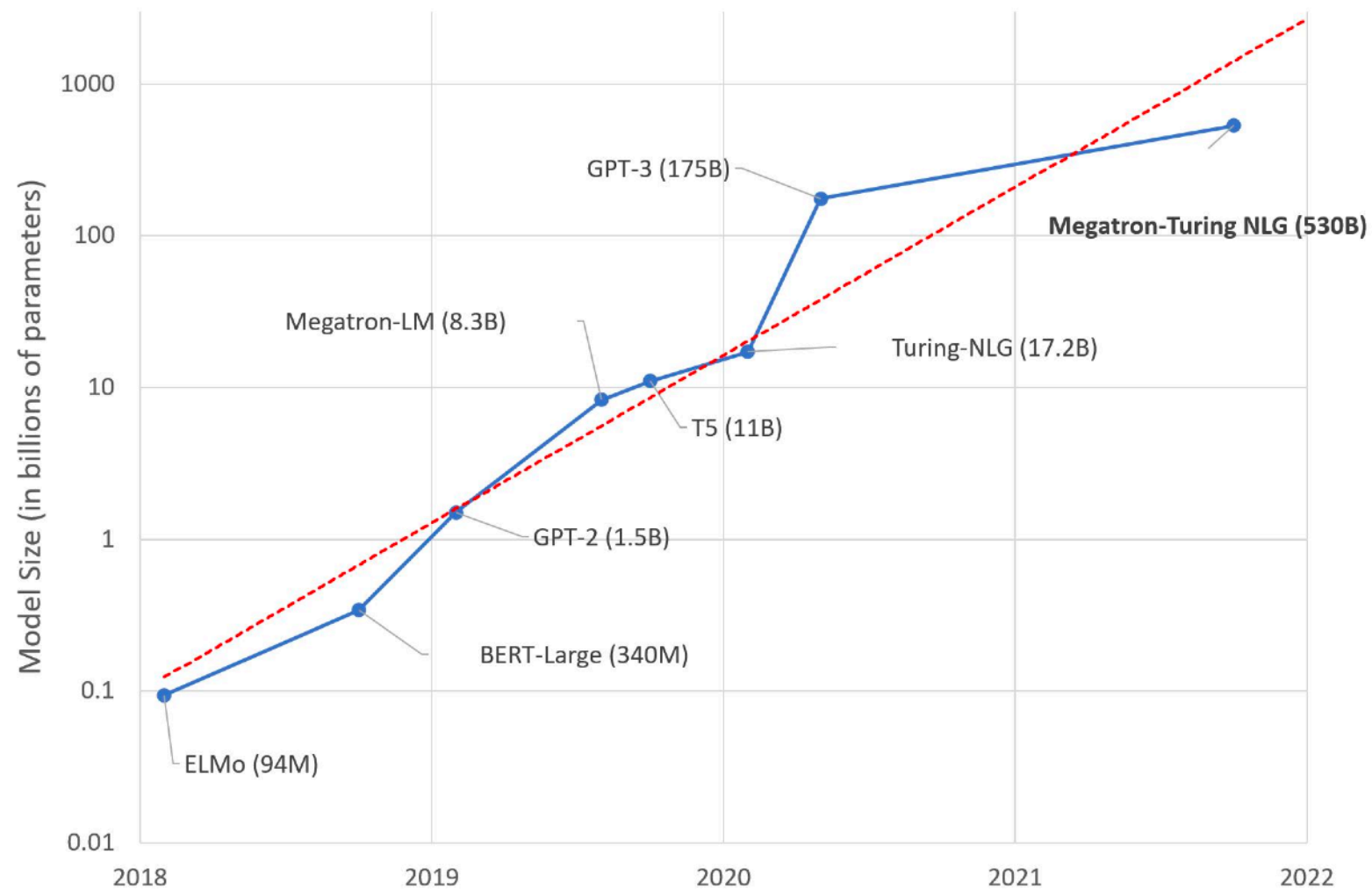


Distributed ML

Ch 5. Splitting the Model

Training Large Model

- Large Model size



- GPU Memory Size

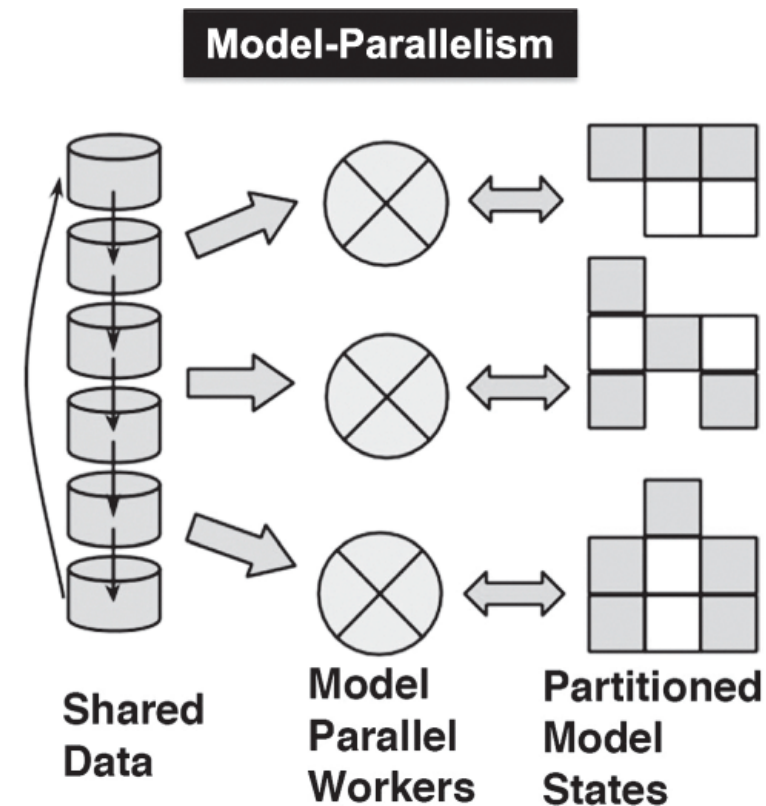
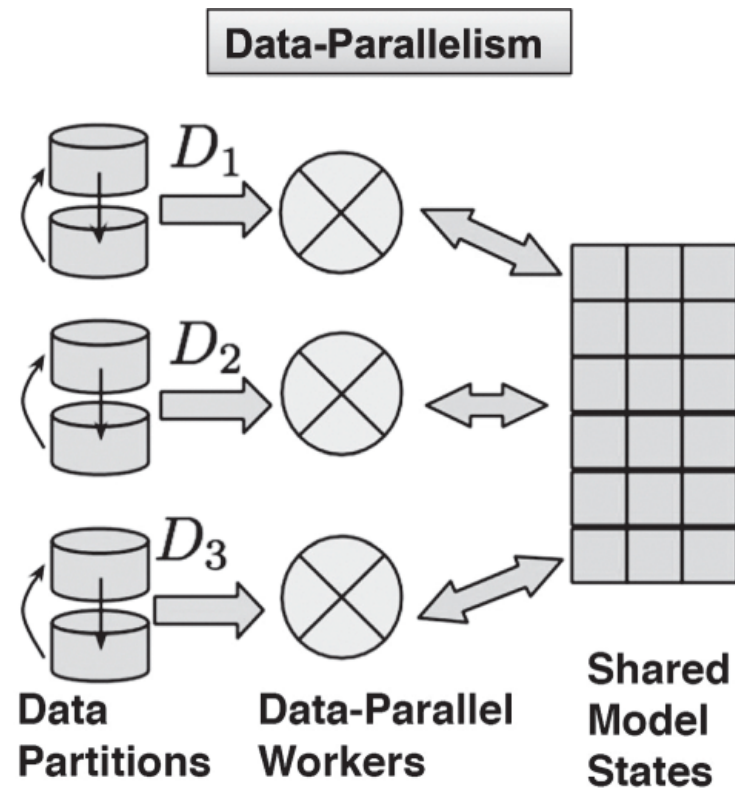
- 16GB (2016) -> 32GB (2017) -> 80GB (2020)

Training Large Model

- Large model training on Single-GPU
 - Out of memory 발생
 - Batch size 를 극단적으로 낮추면 가능
 - Parameter precision 을 극단적으로 낮추면 가능
 - 모델을 split 해서 여러 sequence 로 나누어 연산 수행

-> 학습 시간이 극단적으로 증가

Model Parallel



- Data parallel

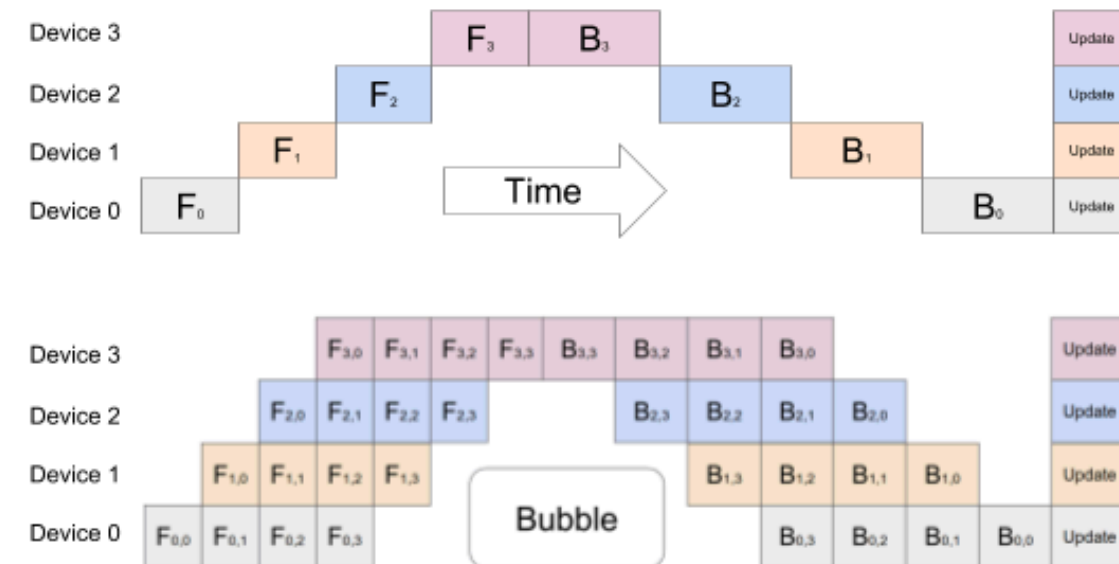
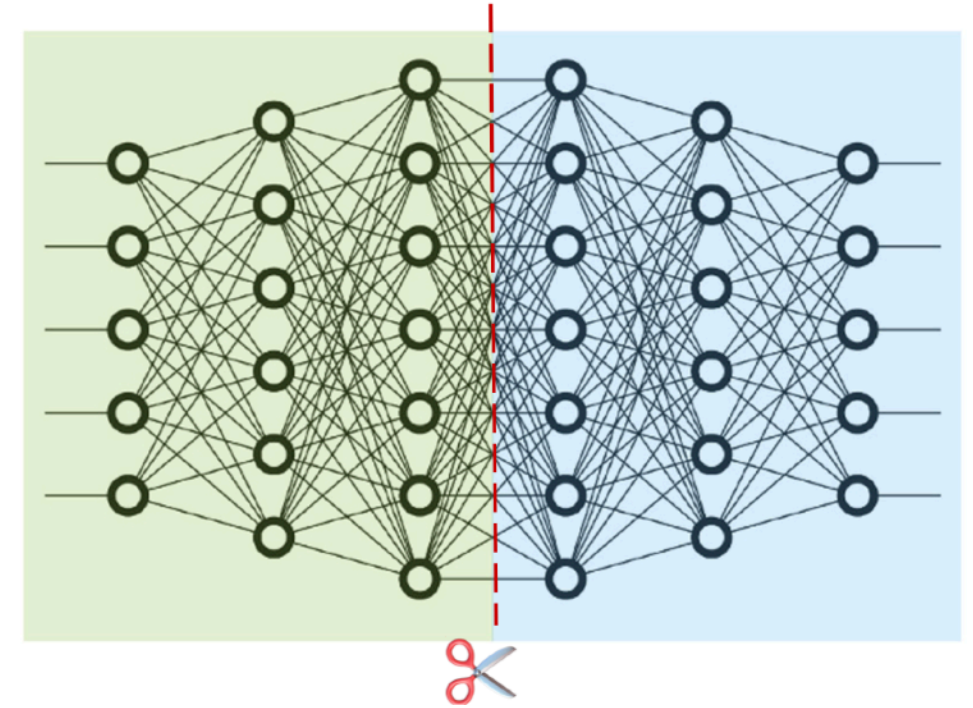
- Dataset 을 분할하여 각 Worker 에 분배
- 각 Worker 간에는 Iter 단위로 N/W 발생
- Grad, weight 와 같이 동일크기의 데이터 교환
- 모든 Worker 는 동일한 용량의 메모리 사용

- Model parallel

- Weight 와 같은 Parameter 분할하여 분배
- 각 Worker 간에는 Iter 도중에 N/W 발생
- 연산 중간 결과 데이터 를 공유하기에 Worker 별 다른 크기의 데이터 교환
- 분할 방식에 따라 각기 다른 용량의 메모리 사용

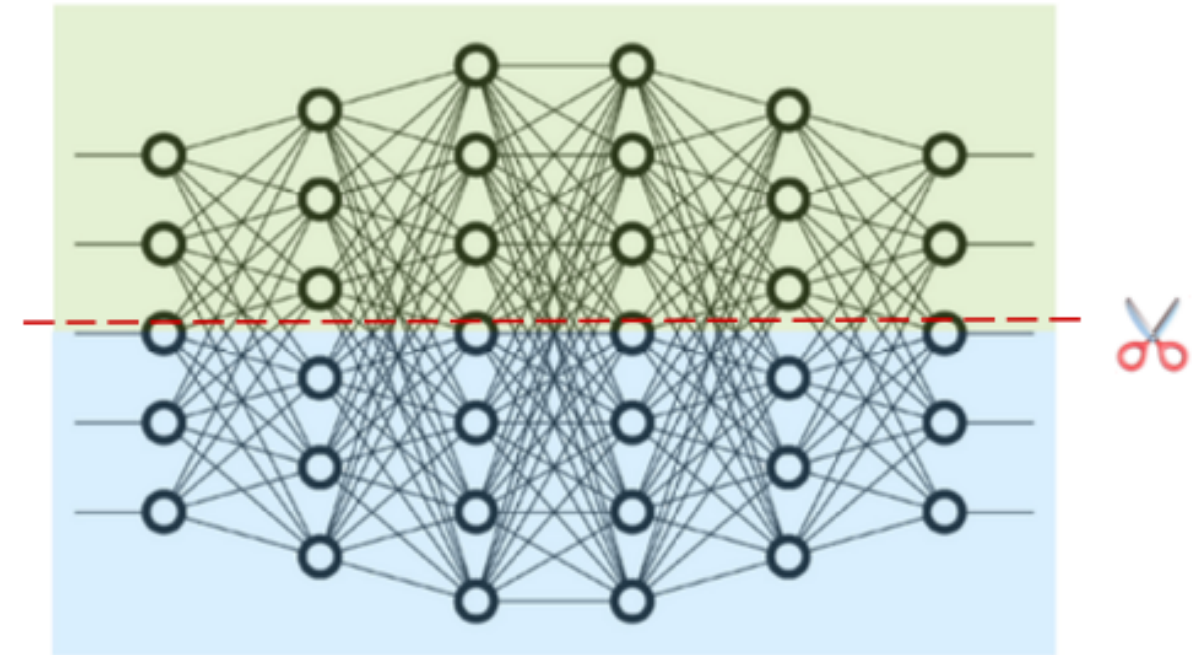
How to split the model

- Inter-layer Parallel (Layer split)
 - 연속된 layer 단위로 분할하여 각 Worker 에 분배
 - 각 Worker 는 중간 결과 값 만을 교환 (N/W 가 단순)
 - 상대적으로 구현이 Simple
 - 완벽한 로드 밸런싱에 취약
 - Bubble 로 인해 GPU 의 Down-time 발생



How to split the model

- Intra-layer Parallel (Tensor Split)
- 개별 layer 내에 Weight 를 분할하여 분배
- 연산 중간에 Aggregation/Split 반복 (Inter-layer 대비 N/W 증가)
- 로드 밸런싱이 용이
- Bubble이 발생하지 않음



$$\begin{array}{|c|c|c|c|} \hline 0 & 1 & 2 & 3 \\ \hline 4 & 5 & 6 & 7 \\ \hline \end{array} \cdot \begin{array}{|c|c|} \hline 10 & 14 \\ \hline 11 & 15 \\ \hline 12 & 16 \\ \hline 13 & 17 \\ \hline \end{array} = \begin{array}{|c|c|} \hline 74 & 98 \\ \hline 258 & 346 \\ \hline \end{array}$$

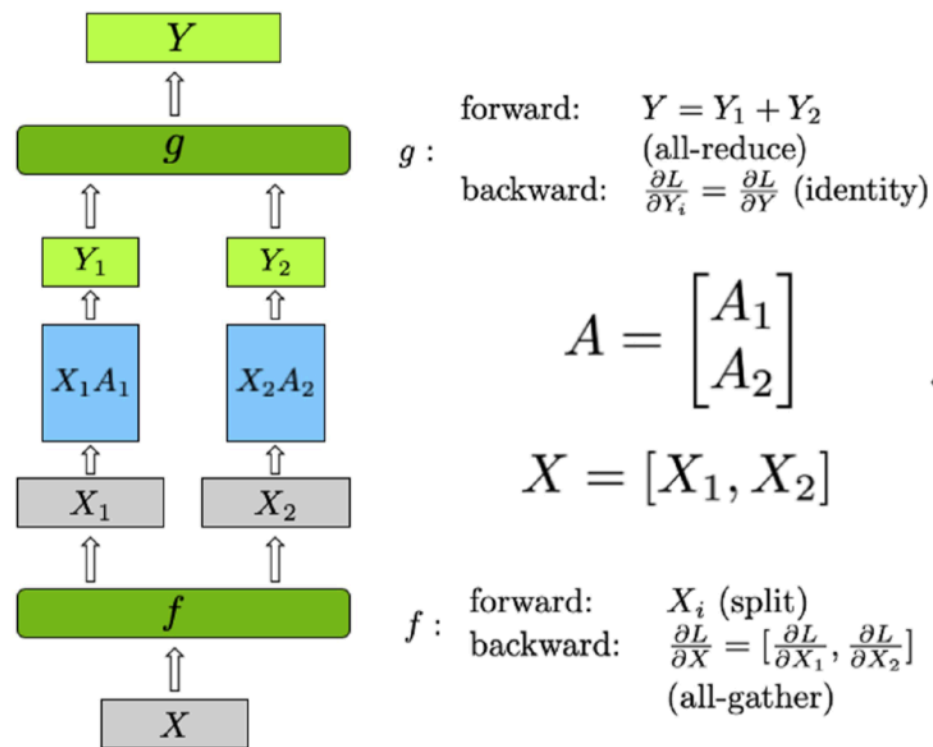
X A Y

$$\begin{array}{ccc} \begin{array}{|c|c|c|c|} \hline 0 & 1 & 2 & 3 \\ \hline 4 & 5 & 6 & 7 \\ \hline \end{array} & \cdot & \begin{array}{|c|} \hline 10 \\ 11 \\ 12 \\ 13 \\ \hline \end{array} \\ X & & A1 \\ \hline \end{array} \quad \begin{array}{ccc} \begin{array}{|c|c|c|c|} \hline 0 & 1 & 2 & 3 \\ \hline 4 & 5 & 6 & 7 \\ \hline \end{array} & \cdot & \begin{array}{|c|} \hline 14 \\ 15 \\ 16 \\ 17 \\ \hline \end{array} \\ X & & A2 \\ \hline \end{array} \quad \begin{array}{ccc} \begin{array}{|c|} \hline 74 \\ 258 \\ \hline \end{array} & = & \begin{array}{|c|} \hline 98 \\ 346 \\ \hline \end{array} \\ Y1 & & Y1 \\ \hline \end{array} \quad \begin{array}{ccc} \begin{array}{|c|c|} \hline 74 & 98 \\ \hline 258 & 346 \\ \hline \end{array} & = & \begin{array}{|c|c|} \hline 74 & 98 \\ \hline 258 & 346 \\ \hline \end{array} \\ \text{cat} & & Y \\ \hline \end{array}$$

$$\begin{array}{ccc} \begin{array}{|c|c|} \hline 0 & 1 \\ \hline 4 & 5 \\ \hline \end{array} & \cdot & \begin{array}{|c|c|} \hline 10 & 14 \\ \hline 11 & 15 \\ \hline \end{array} \\ X1 & & A1 \\ \hline \end{array} \quad \begin{array}{ccc} \begin{array}{|c|c|} \hline 2 & 3 \\ \hline 6 & 7 \\ \hline \end{array} & \cdot & \begin{array}{|c|c|} \hline 12 & 16 \\ \hline 13 & 17 \\ \hline \end{array} \\ X2 & & A2 \\ \hline \end{array} \quad \begin{array}{ccc} \begin{array}{|c|c|} \hline 11 & 15 \\ \hline 95 & 131 \\ \hline \end{array} & = & \begin{array}{|c|c|} \hline 63 & 83 \\ \hline 163 & 215 \\ \hline \end{array} \\ Y1 & & Y2 \\ \hline \end{array} \quad \begin{array}{ccc} \begin{array}{|c|c|} \hline 74 & 98 \\ \hline 258 & 346 \\ \hline \end{array} & = & \begin{array}{|c|c|} \hline 74 & 98 \\ \hline 258 & 346 \\ \hline \end{array} \\ + & & Y \\ \hline \end{array}$$

Tensor Split

Row-wise split



Column wise split

