# ZeRO
# (Zero Redundancy Optimizer)

# 목 차

# Problem to solve

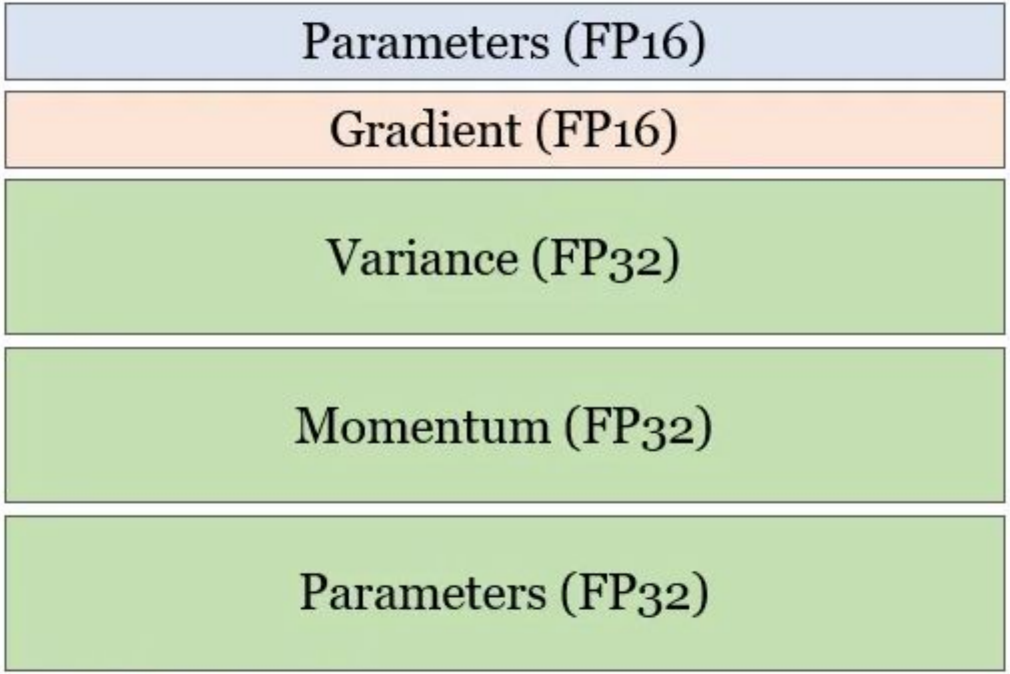
- Large model을 학습하는 다른 방법에 불필요한 redundancy가 있음

# Points to learn

- transformer model의 memory(training) 요구량 계산 방법
- DP,MP,PP에서 redundant memory 사용과 비효율성
- 이 논문에서 제안하는 ZeRO-DP, ZeRO-R

# Background

- Training시 memory는
  - model state
  - residual state로 나뉨
- model state는
  - model parameter
  - gradient
  - optimizer state
- residual state는
  - activation
  - temporary buffer
  - unusable memory fragments



# Background(Model state memory)

- parameter와 gradient는 16bit으로 각각 n_params * 2byte의 size를 가짐
- Optimizer state는 n_params * 12byte의 size를 가짐
    - momentum(time averaged momentum) - 32bit
    - variance(variance of the gradients) - 32bit
    - parameters - 32bit

# Background(Residual state memory)

Activation

- seq len 1K, 32 batch, 1.5B model에 대해서 60GB정도 필요함
    - activation recomputation을 통해 약 8GB정도로 줄일 수 있음

Temporary buffers

- allreduce와 같은 comm연산 할 때 buffer에 flatten할 때 필요한 메모리
- 1.5B model에 대해서 32bit buffer 할당 시 6GB정도 필요함

Memory fragmentation

- contiguous memory가 부족한 현상.
- large model training시 30%정도의 memory 여유가 있음에도 불구하고 out of memory 발생

# Background(DP,MP,PP 단점)

DP

- model state를 모든 GPU에 copy해야 함. redundant

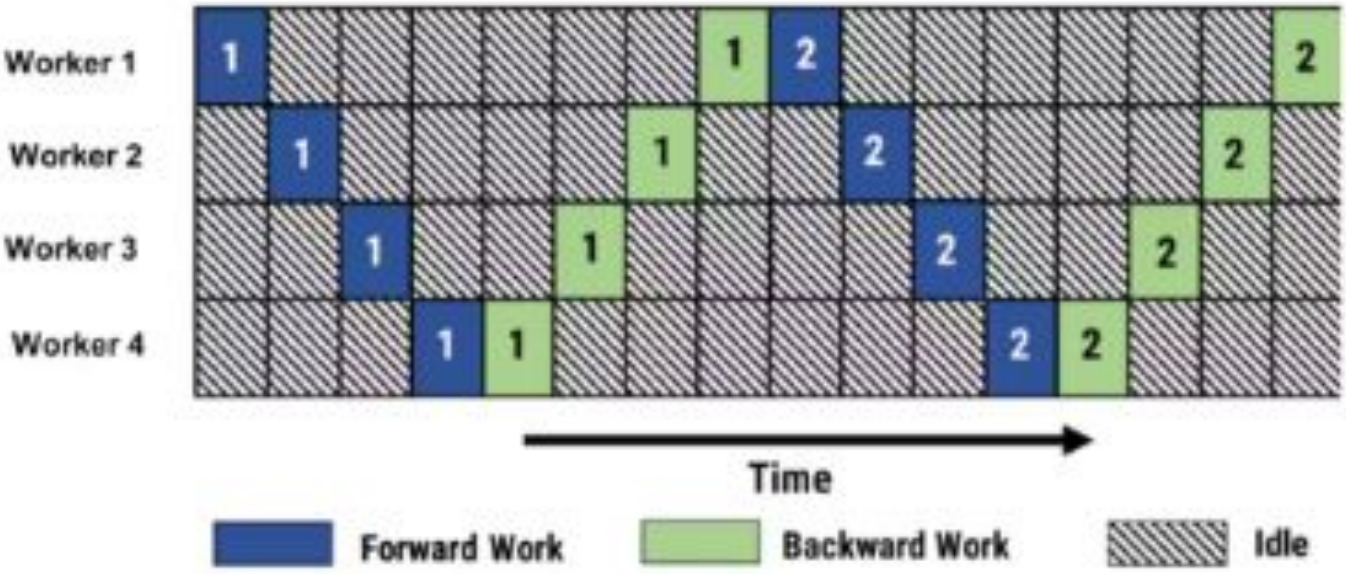
MP

- model을 vertically 쪼개는거라서 communication overhead가 크고, 연구자들이 코드로 적용하기 많이 어려움

PP

- bubble을 없애기 위해서 DP를 충분히 추가해줘야 함.

# Background(PP의 bubble)




Model Parallelism

# Solution(ZeRO-DP)

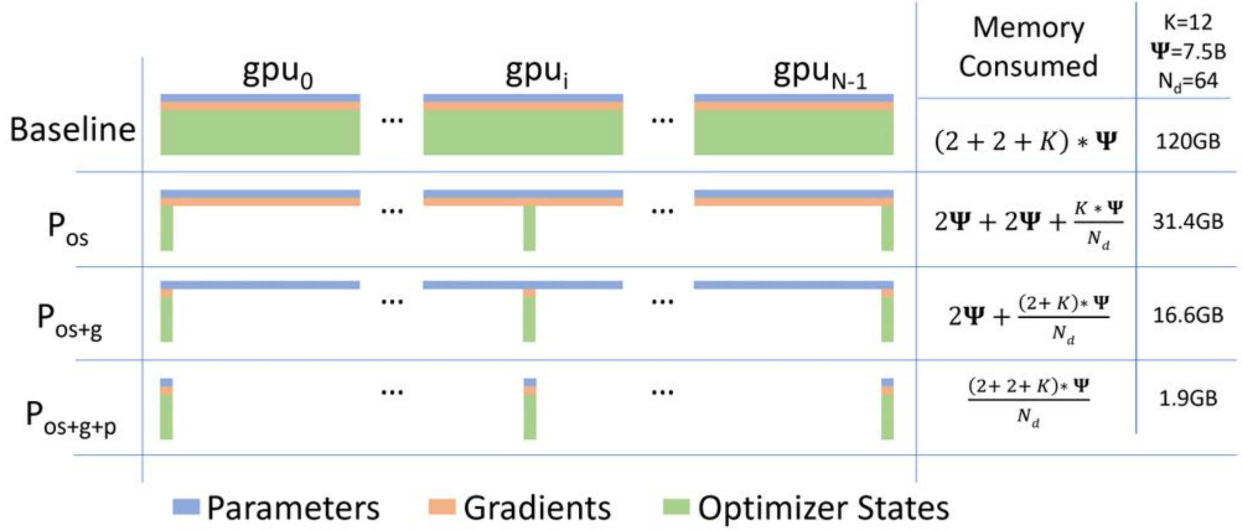
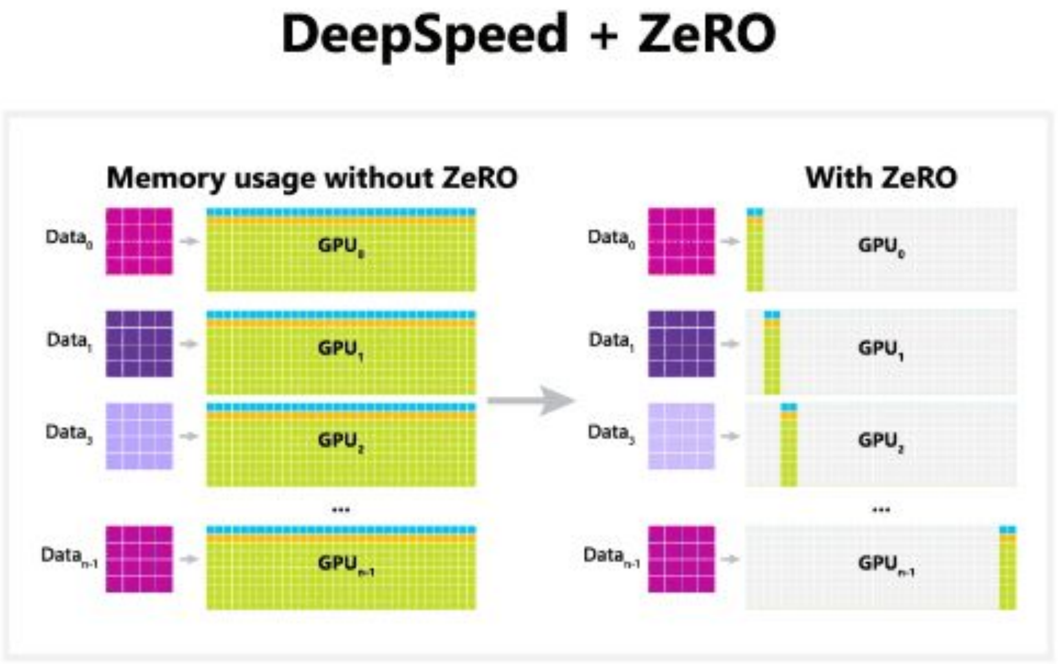



Figure 1: Comparing the per-device memory consumption of model states, with three stages of *ZeRO*-DP optimizations. $\Psi$ denotes model size (number of parameters), $K$ denotes the memory multiplier of optimizer states, and $N_d$ denotes DP degree. In the example, we assume a model size of $\Psi = 7.5B$ and DP of $N_d = 64$ with $K = 12$ based on mixed-precision training with Adam optimizer.

# Solution(ZeRO-DP)



# Experiments

| DP | 7.5B Model (GB) | | | 128B Model (GB) | | | 1T Model (GB) | | |
|---|---|---|---|---|---|---|---|---|---|
| | $P_{os}$ | $P_{os+g}$ | $P_{os+g+p}$ | $P_{os}$ | $P_{os+g}$ | $P_{os+g+p}$ | $P_{os}$ | $P_{os+g}$ | $P_{os+g+p}$ |
| 1 | 120 | 120 | 120 | 2048 | 2048 | 2048 | 16000 | 16000 | 16000 |
| 4 | 52.5 | 41.3 | **30** | 896 | 704 | 512 | 7000 | 5500 | 4000 |
| 16 | 35.6 | **21.6** | 7.5 | 608 | 368 | 128 | 4750 | 2875 | 1000 |
| 64 | **31.4** | 16.6 | 1.88 | 536 | 284 | **32** | 4187 | 2218 | 250 |
| 256 | 30.4 | 15.4 | 0.47 | 518 | 263 | 8 | 4046 | 2054 | 62.5 |
| 1024 | 30.1 | 15.1 | 0.12 | 513 | 257 | 2 | 4011 | 2013 | **15.6** |

Table 1: Per-device memory consumption of different optimizations in *ZeRO*-DP as a function of DP degree . Bold-faced text are the combinations for which the model can fit into a cluster of 32GB V100 GPUs.

# Results

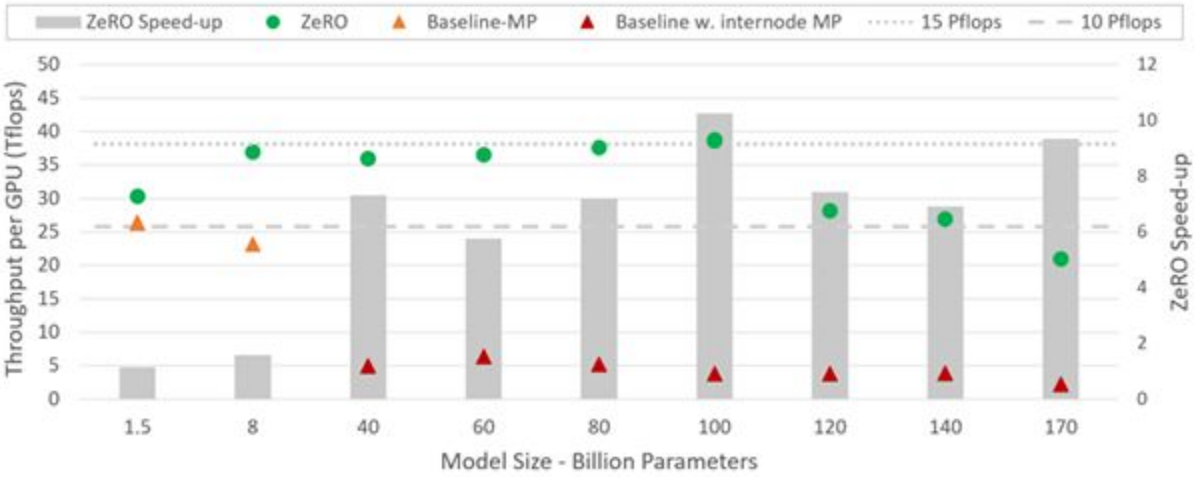



Figure 2: *ZeRO* training throughput and speedup w.r.t SOTA baseline for varying model sizes. For *ZeRO*, the MP always fit in a node, while for baseline, models larger than 40B require MP across nodes.

# Results

| MP | GPUs | Max Theoretical Model Size | | | | Measured Model Size | |
|---|---|---|---|---|---|---|---|
| | | Baseline | $P_{os}$ | $P_{os+g}$ | $P_{os+g+p}$ | Baseline | *ZeRO*-DP ($P_{os}$) |
| 1 | 64 | 2B | **7.6B** | 14.4B | 128B | 1.3B | **6.2B** |
| 2 | 128 | 4B | **15.2B** | 28.8B | 256B | 2.5B | **12.5B** |
| 4 | 256 | 8B | **30.4B** | 57.6B | 0.5T | 5B | **25B** |
| 8 | 512 | 16B | **60.8B** | 115.2B | 1T | *10B* | **50B** |
| 16 | 1024 | 32B | **121.6B** | 230.4B | *2T* | 20B | **100B** |

Table 2: Maximum model size through memory analysis (left) and the measured model size when running with *ZeRO-OS* (right). The measured model size with $P_{os}$ matches the theoretical maximum, demonstrating that our memory analysis provides realistic upper bounds on model sizes.