DeepSpeed

Extreme-scale model training for everyone

HUMANE Lab

김태균

25.01.24

Overview

- What is DeepSpeed?
- Four core technologies
- DeepSpeed in NLP

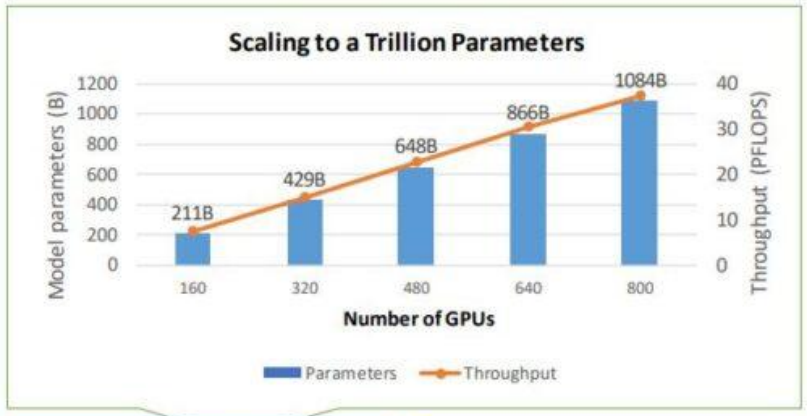
What is DeepSpeed?

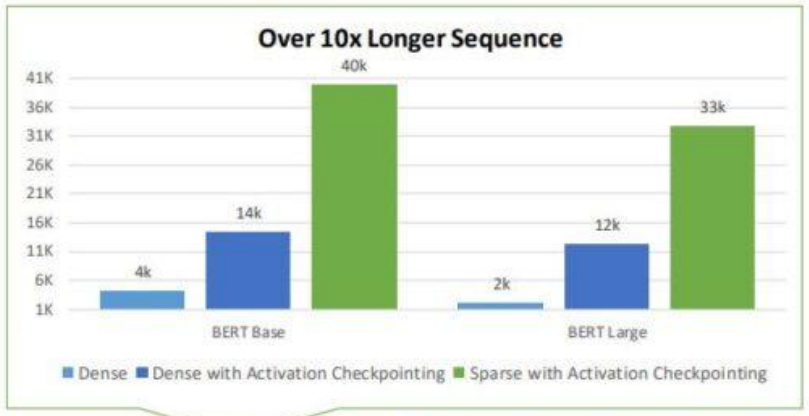
DeepSpeed

: Open-source deep learning optimization library developed by Microsoft

- It is designed to efficiently train and run large-scale models
  - Scale expansion
  - Speed improvement
  - Cost reduction
  - Accessibility enhancement

Four Core Technologies



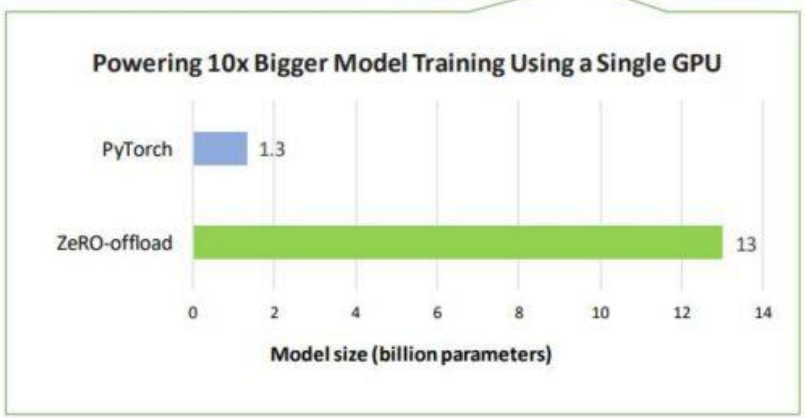


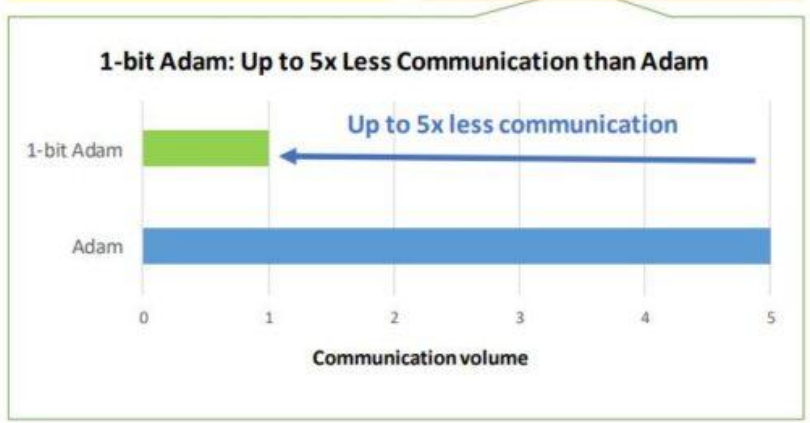
3D Parallelism
- 1 trillion parameter model training

ZeRO-Offload
- 13B model on single GPU, 10x bigger

Sparse Attention
- 10x longer sequence, up to 6x faster

1-bit Adam
- 5x less communication





3D Parallelism

- Combined three powerful technologies to enable training trillion-scale models
  - Data parallelism
  - Model parallelism
  - Pipeline parallelism

⇒ Improving memory and compute efficiency

3D Parallelism

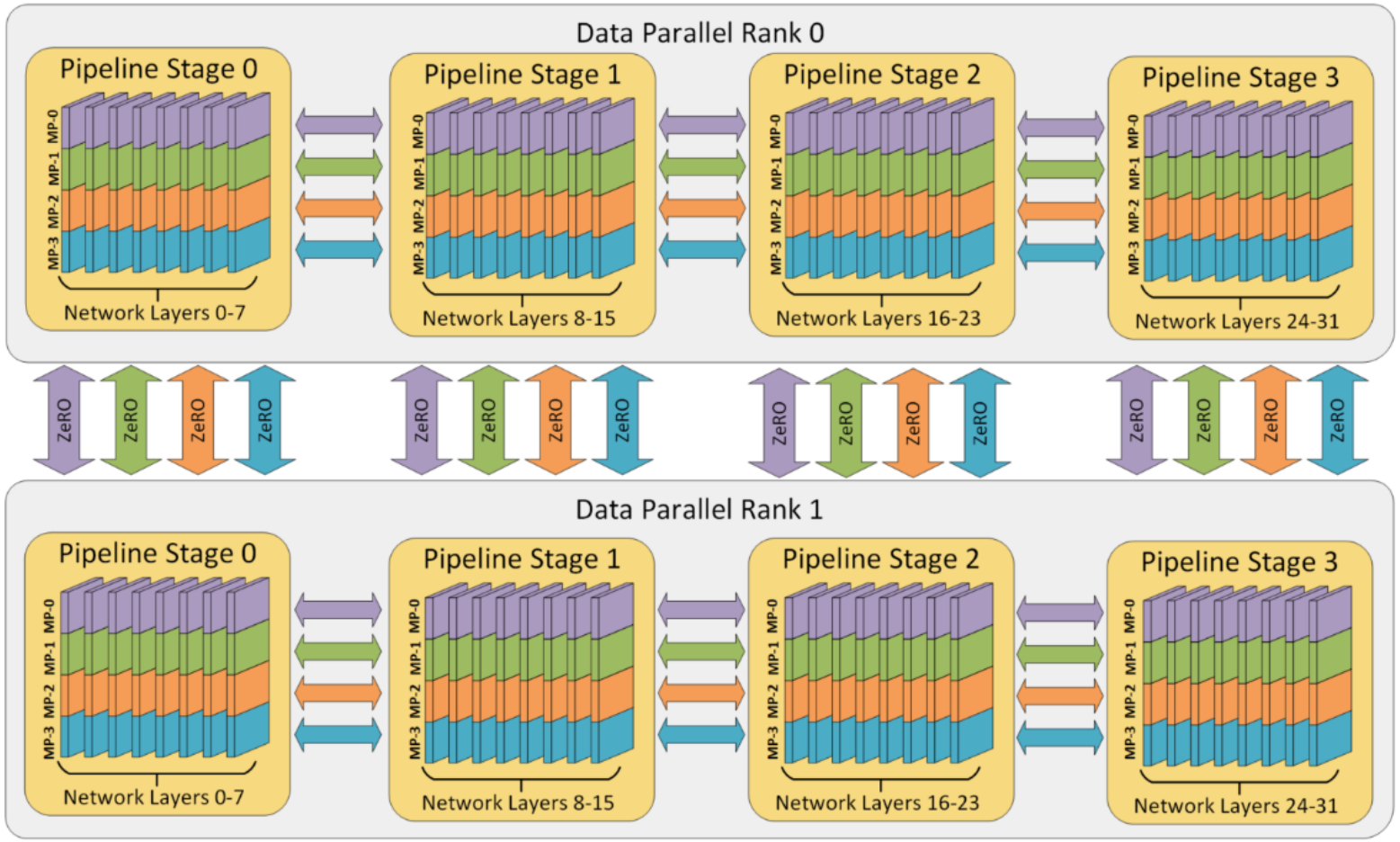


Figure 1: Example 3D parallelism with 32 workers. Layers of the neural network are divided among four pipeline stages. Layers within each pipeline stage are further partitioned among four model parallel workers. Lastly, each pipeline is replicated across two data parallel instances, and ZeRO partitions the optimizer states across the data parallel replicas.

3D Parallelism

- Topology aware 3D mapping

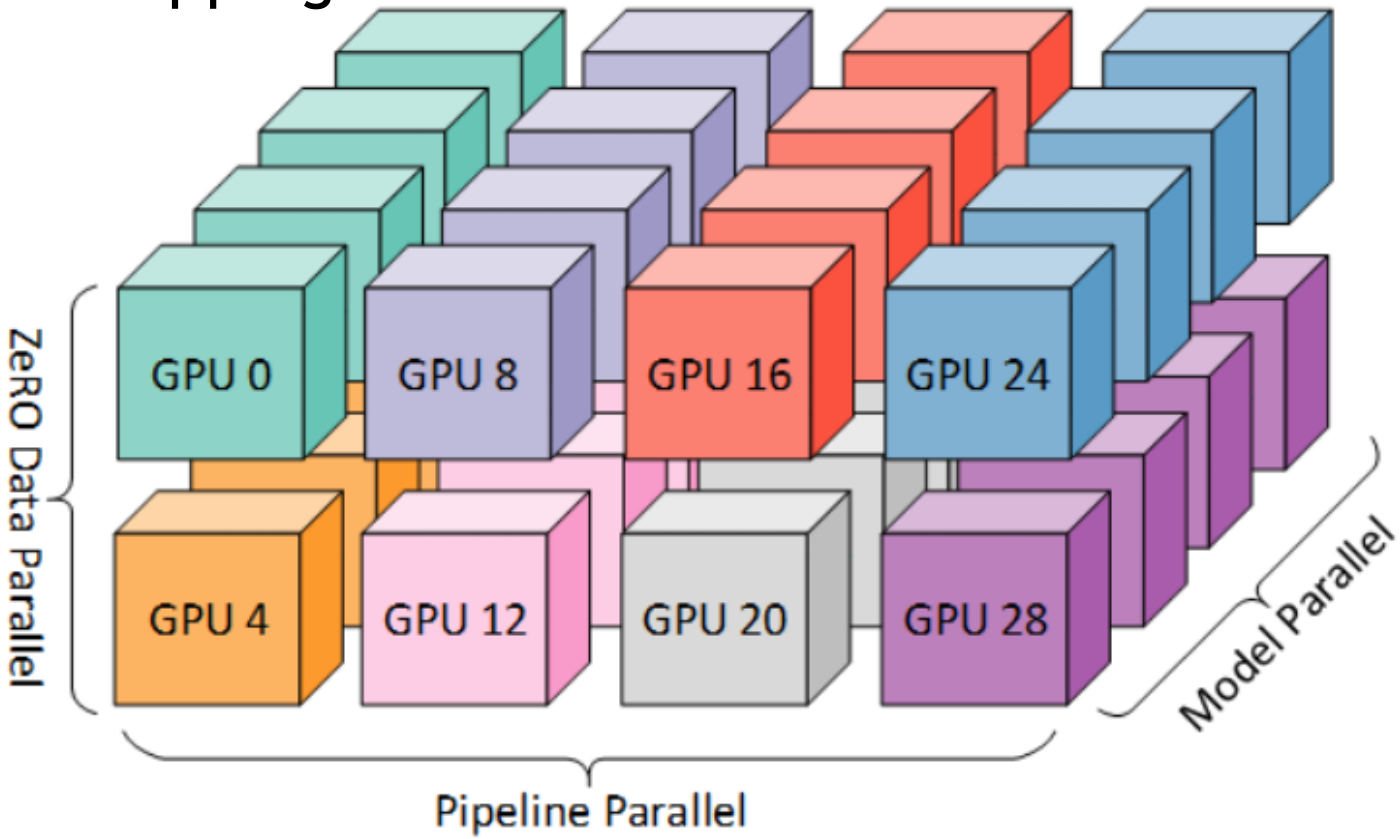


Figure 2: Mapping of workers in Figure 1 to GPUs on a system with eight nodes, each with four GPUs. Coloring denotes GPUs on the same node.

ZeRO-Offload

- By enabling multi-billion-parameter model training on a single GPU
- Offloads optimizer states and gradients to CPU memory, built on ZeRO-2

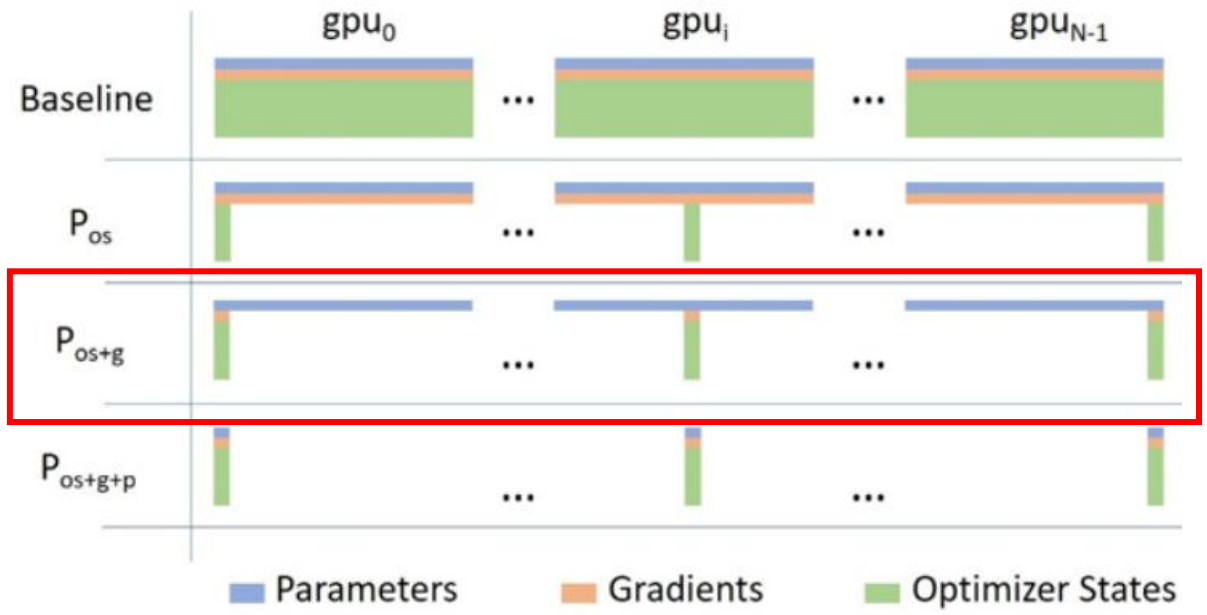
⇒ Enables deep learning practitioners with limited resources to access large-scale model training

ZeRO-Offload

ZeRO (Zero Redundancy Optimizer)

- Optimization technology designed to maximize memory efficiency

ZeRO-2



Sparse Attention

- Attention-based deep learning models are highly effective in capturing relationships between tokens in an input sequence
- However, their application to long sequence input is limited by compute and memory requirements of the attention computation that grow quadratically

Sparse attention kernels

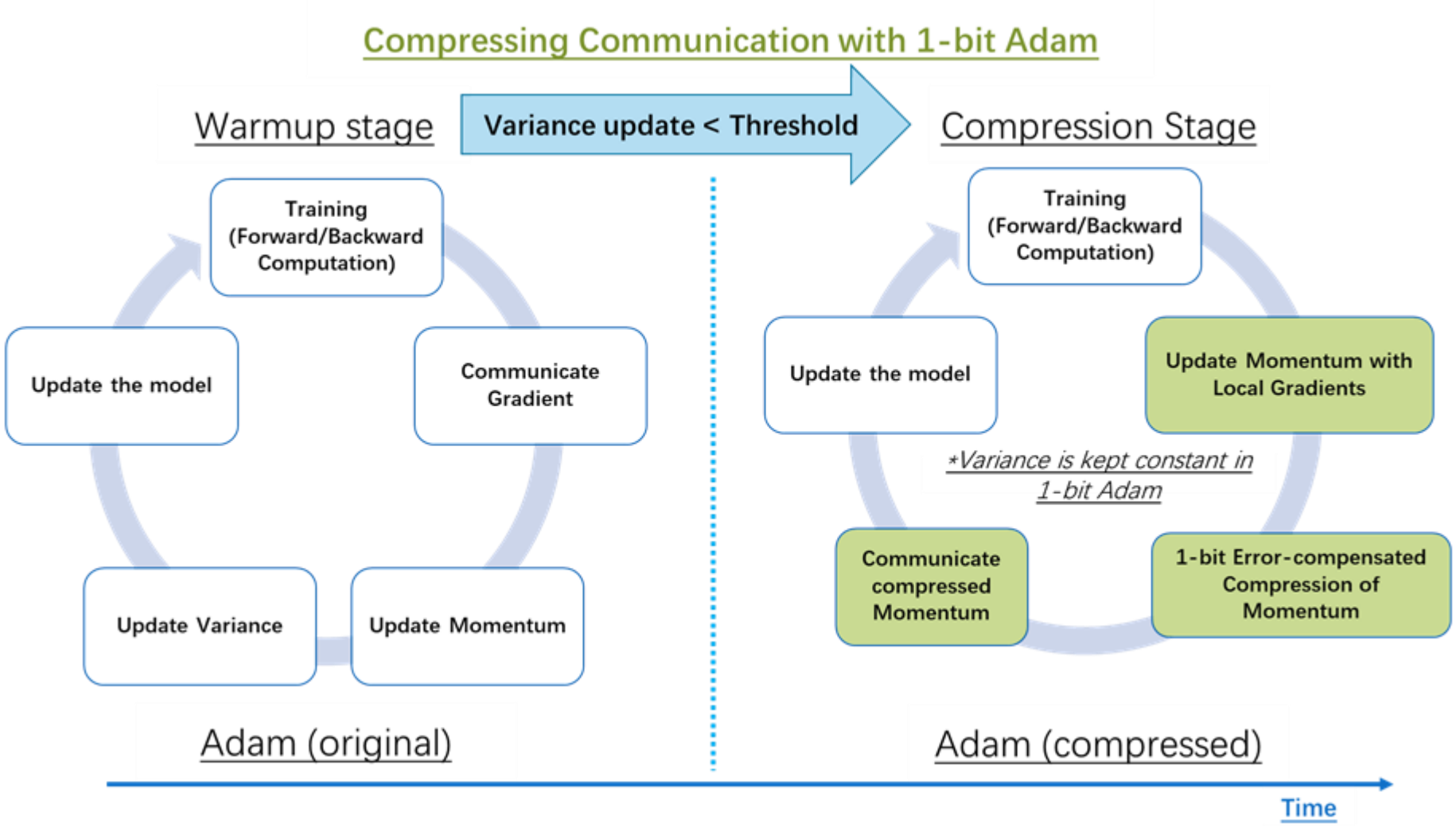
⇒ Reduce the compute and memory requirement of attention computation by orders of magnitude via block-sparse computation

1-bit Adam

- Scalable training of large models requires careful optimization
- From a system standpoint, communication has become a major bottleneck
- Communication compression is an important technique to reduce training time on such systems
- For a powerful optimizer like Adam, the non-linear dependency on gradient makes it challenging to develop compression techniques

⇒ Compressing communication with 1-bit Adam

1-bit Adam


Compressing Communication with 1-bit Adam
Warmup stage
Variance update < Threshold
Compression Stage
Training (Forward/Backward Computation)
Communicate Gradient
Update Momentum
Update Variance
Update the model
Adam (original)
Training (Forward/Backward Computation)
Update Momentum with Local Gradients
1-bit Error-compensated Compression of Momentum
Communicate compressed Momentum
Update the model
*Variance is kept constant in 1-bit Adam
Adam (compressed)
Time

DeepSpeed in NLP

config.json > ...

```
{
    "train_batch_size": 16,
    "gradient_accumulation_steps": 2,
    "fp16": {
      "enabled": true
    },
    "zero_optimization": {
      "stage": 2,
      "offload_optimizer": {
        "device": "cpu",
        "pin_memory": true
      }
    }
}
```

```
import deepspeed

model_engine, optimizer, _, _ = deepspeed.initialize(
    model=model,
    config=ds_config,
    model_parameters=model.parameters()
)


for epoch in range(3):
    outputs = model_engine(**inputs)
    loss = outputs.loss
    model_engine.backward(loss)
    model_engine.step()
```

Conclusion

DeepSpeed is deep learning library that enables the efficient training and execution of extreme-scale models