
大模型：分布式训练
DeepSpeed
ZOMI

大模型业务全流程

数据 & 模型算法

2. 数据处理

- 开源数据
- 数据预处理
- 向量数据库

3. 模型算法

- LLM模型架构
- 多模态一切皆Tokens

模型训练 & 微调

4. 模型训练

- 混合精度
- 梯度检查
- 梯度累积
- …
- 5. 分布式并行
- 训练集群稳定性

6. 模型微调

- 全参微调
- 低参微调
- 指令微调

模型验证 & 推理部署

7. 模型验证

- 下游任务
- 测评标准
- bench mark

8. 推理与智能体

- 量化压缩
- 推理加速
- 9. Agent 智能体

集群算力准备

1. AI 集群建设

- 计算、存储、网络
- AI 集群机房建设
- AI 集群上线与运维

大模型系列 – 分布式训练加速

- **具体内容**
  - **分布式加速库：**
    - 业界常用分布式加速库 & 作用
  - **DeepSpeed 特性：**
    - 基本概念 - 整体框架 – Zero-1/2/3 – ZeRO-Offload – ZeRO-Infinity
  - **Megatron 特性：**
    - 总体介绍 – 整体流程 – 并行配置 – DP – TP – PP

2. DeepSpeed

https://github.com/microsoft/DeepSpeed

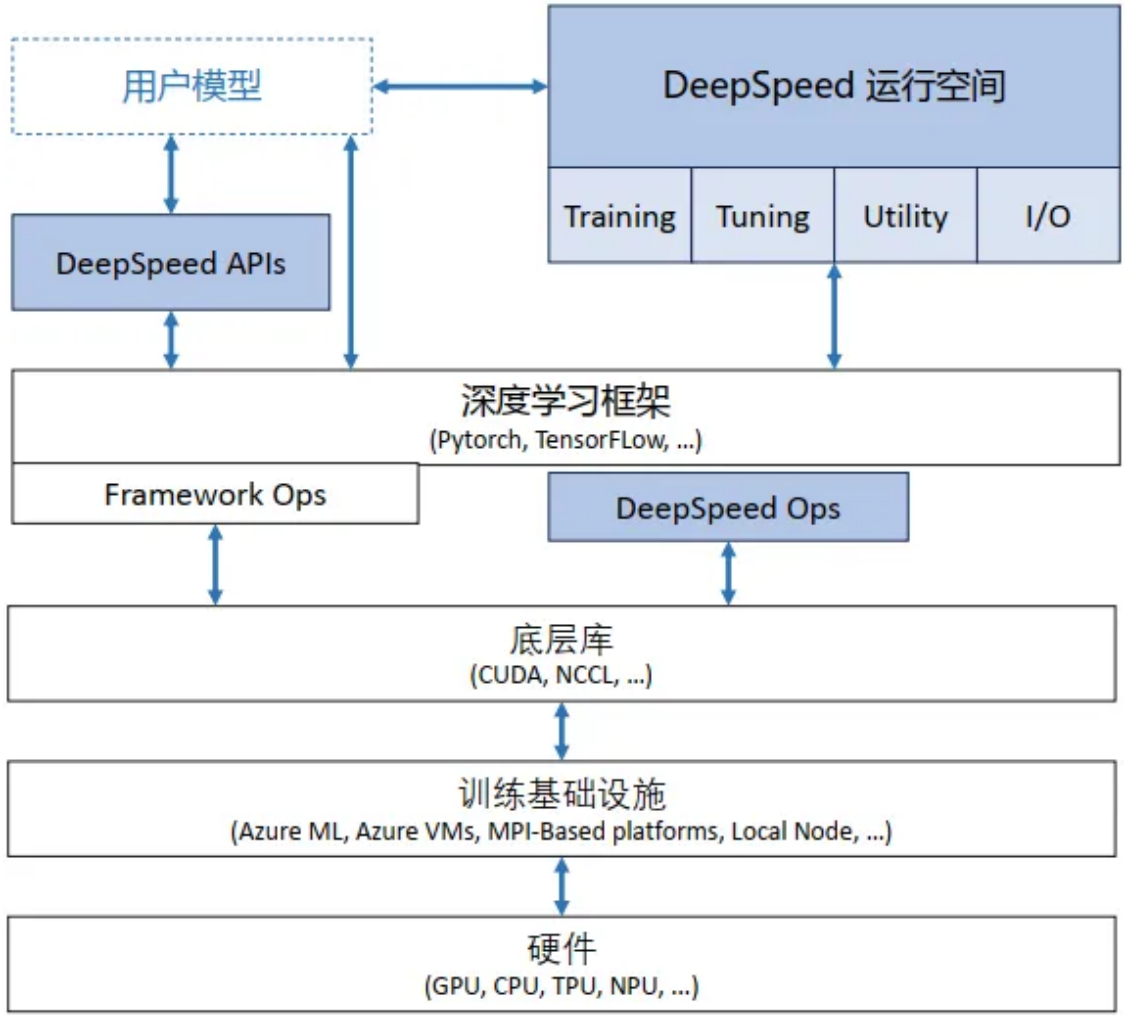
创新特性

- **Training**：为大模型训练供 ZeRO、3D-Parallelism、DeepSpeed-MoE、ZeRO-Infinity等特性；
- **Inference**：提供Tensor、Pipeline、Expert等并行特性与推理内核、通信优化和异构内存特性结合；
- **Compression**：提供 ZeroQuant 和 XTC 等 SoTA 在压缩方面的创新；
- **4Science**：遥远的你；

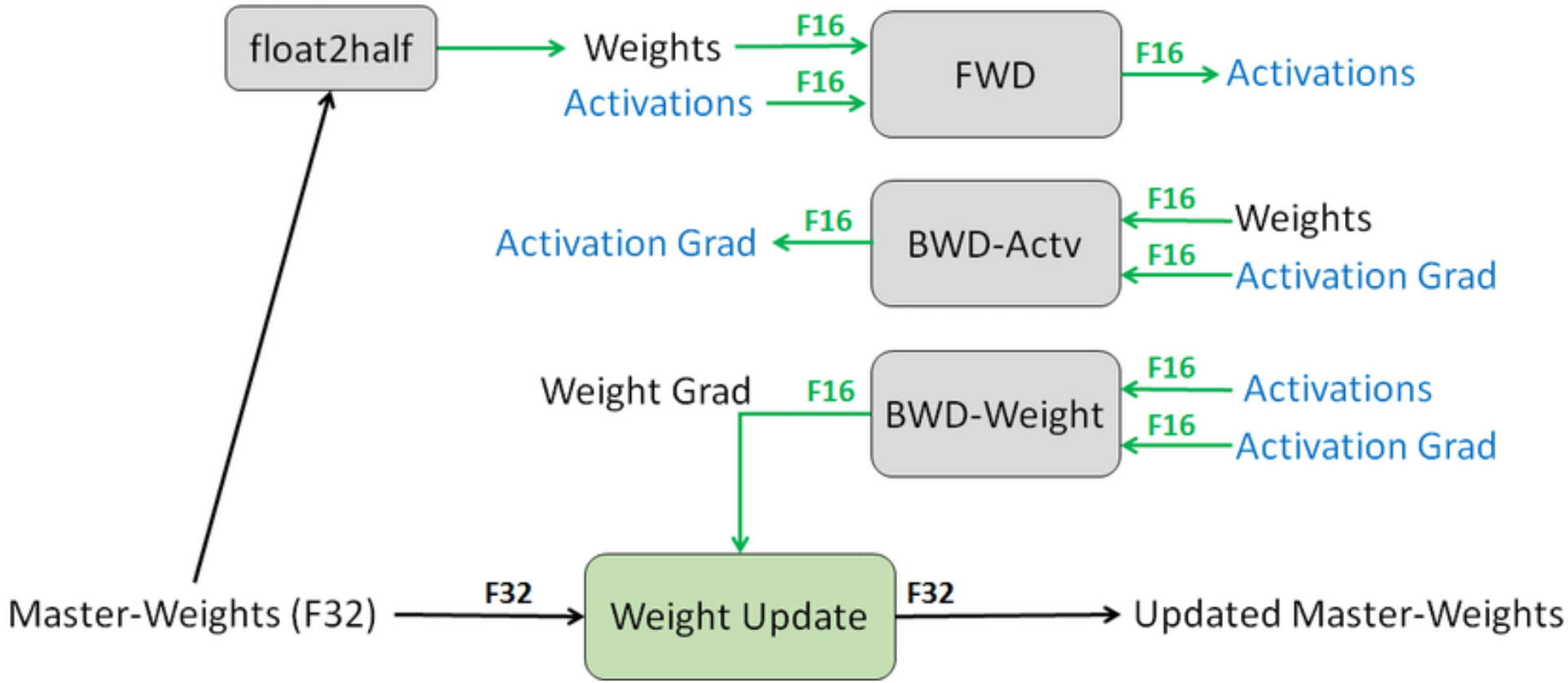
Training	Inference	Compression	Science
• Speed Scale Cost • Democratization • MoE models • Long sequence • RLHF	• Large models • Latency • Serving cost • Agility	• Model size • Latency • Composability • Runnable on client devices	• Speed • Scale • Capability • Diversity • Discovery

软件架构

1. APIs：
   - 配置参数在 ds_config.json 中，通过 API 接口可以调用 DeepSpeed 训练/推理模型；
2. RunTime：
   - 核心运行时组件，负责管理、执行和优化性能。包括数据、模型、并行优化、微调、故障检测以及 CheckPoint 保存和加载等任务。
3. Ops：
   - 底层内核组件，使用C++和CUDA实现。优化计算和通信，提供底层操作；



混合精度训练 && 显存占用分析



混合精度训练 && 显存占用分析

- Model States 模型本身相关且必须存储的参数：
  1. Parameters：模型参数
  2. Gradients：模型梯度
  3. Optimizer States：Adam中 momentum和variance
- Residual States 非模型必须，训练过程中产生的参数：
  1. Activation：激活值
  2. Temporary Buffers：临时存储
  3. Unusable Fragmented Memory：碎片化存储空间

ZeRO 并行

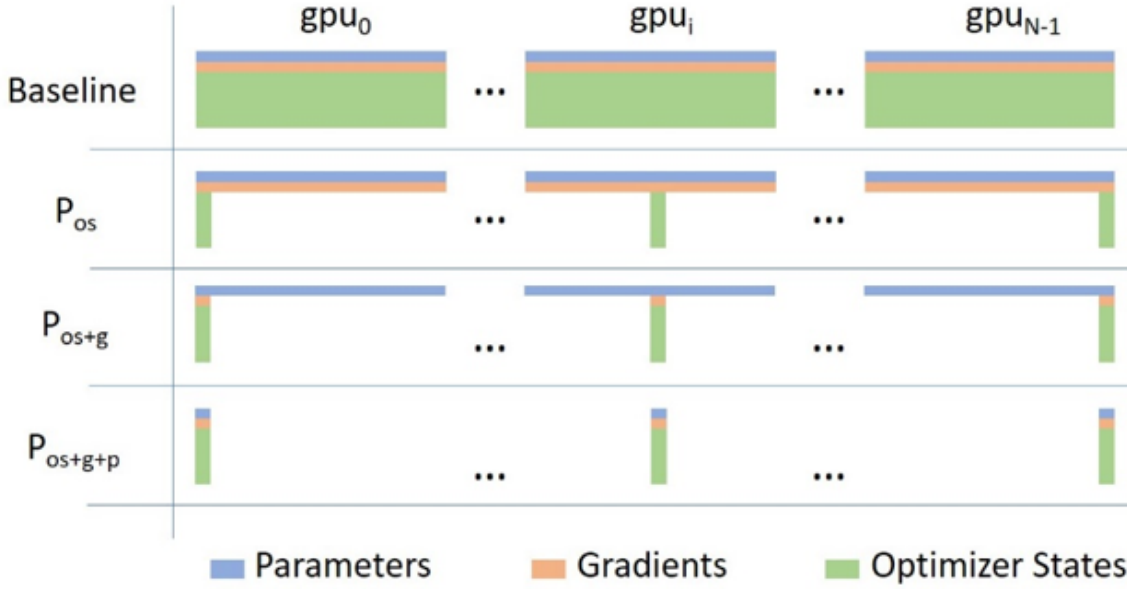
- ZeRO Zero Redundancy Optimizer，一系列显存优化方法的统称：
  1. ZeRO-DP（Data Parallel）：ZeRO1/2/3
  2. ZeRO-R（Reduce）： Activation Checkpointing、 Memory Defragmentation
  3. ZeRO-Offload： Offload Strategy && Offload Schedule
  4. ZeRO-Infinity： Breaking the GPU Memory Wall for Extreme Scale Deep Learning

ZeRO-DP (Data Parallel)

- ZeRO Zero Redundancy Optimizer，一系列显存优化方法的统称：
  1. ZeRO-DP (Data Parallel) ：ZeRO1/2/3
- Optimizer state partitioning (ZeRO stage 1) ：
  - 只对 optimizer 状态进行切分，占用内存原始1/4；
- Gradient partitioning (ZeRO stage 2) ：
  - 对 optimizer 和 grad 进行切分，占用内存原始1/8；
- Parameter partitioning (ZeRO stage 3) ：
  - 对 optimizer、grad 和模型参数进行切分，内存减少与数据并行度和复杂度成线性关系，同时通信容量是数据并行性的 1.5 倍；

ZeRO-DP (Data Parallel)

- Optimizer state partitioning (ZeRO stage 1) ：只对 optimizer 状态进行切分，占用内存原始1/4；
- Gradient partitioning (ZeRO stage 2) ：对 optimizer 和 grad 进行切分，占用内存原始1/8；
- Parameter partitioning (ZeRO stage 3) ：对 optimizer、grad 和模型参数进行切分，内存与数据并行线性关系，通信 DP 1.5X；



	Memory Consumption		Comm Volume
	Formulation	Specific Example K=12 Ψ=7.5B N_d=64	
Baseline	$(2+2+K)*\Psi$	120GB	1x
P_{os}	$2\Psi+2\Psi+\frac{K*\Psi}{N_d}$	31.4GB	1x
P_{os+g}	$2\Psi+\frac{(2+K)*\Psi}{N_d}$	16.6GB	1x
P_{os+g+p}	$\frac{(2+2+K)*\Psi}{N_d}$	1.9GB	1.5x

ZeRO: Where the memory goes?

- **模型本身相关且必须存储的参数**
- Model States ：
  1. Parameters (half)：2 bytes
  2. Gradients (half)：2 bytes
- Optimizer States ：
  1. Master Weight (fp32)：4 bytes
  2. Adam momentum (fp32)：4 bytes
  3. Adam Variance (fp32)： 4 bytes

- 假设模型参数量 Ψ ，使用Adam优化器混合精度训练：
  - a.模型参数和梯度 float16，显存消耗为 2Ψ + 2Ψ；
  - Adam 维护 float32 模型副本，消耗 4Ψ；
  - Adam 辅助变量 fp32 momentum + fp32 variance，显存消耗 4Ψ+4Ψ；
- 模型消耗 2Ψ+2Ψ=4Ψ ，Adam 消耗 4Ψ+4Ψ+4Ψ=12Ψ，总消耗 Ψ+12Ψ= 16Ψ
- 优化器显存占用表示 KΨ（不同的优化器不同），则、混合精度训练显存占用 4Ψ+KΨ

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem